

Selling to Overconfident Consumers

By MICHAEL D. GRUBB*

Consumers may overestimate the precision of their demand forecasts. This overconfidence creates an incentive for both monopolists and competitive firms to offer tariffs with included quantities at zero marginal cost, followed by steep marginal charges. This matches observed cellular phone service pricing plans in the United States and elsewhere. An alternative explanation with common priors can be ruled out in favor of overconfidence based on observed customer usage patterns for a major US cellular phone service provider. The model can be reinterpreted to explain the use of flat rates and late fees in rental markets, and teaser rates on loans. Nevertheless, firms may benefit from consumers losing their overconfidence. (JEL D12, L11, L96)

Firms commonly offer three-part tariffs, or menus of three-part tariffs, in a variety of contexts. A three-part tariff consists of a fixed fee, an included allowance of units for which the marginal price is zero, and a positive marginal price for additional usage beyond the allowance. A prime example is the US cellular phone services market in which firms typically offer consumers plans consisting of a fixed monthly fee, an allowance of minutes, and an overage rate for minutes beyond the allowance. Pricing of Internet service is similar in many European countries, where usage is billed per megabyte (Anja Lambrecht and Bernd Skiera 2006). Other examples of three-part tariffs include car leasing contracts, which often include a mileage allowance and charge per mile thereafter.¹ In a variety of rental markets, contracts charge a flat rate for a specified period followed by steep late fees. Finally, introductory credit card offers may also be three-part tariffs. For instance a \$1,000 balance might be charged an initial balance transfer fee, zero marginal charge per month for the first six months, and a high marginal charge per month thereafter.

The existing literature on nonlinear pricing does not provide a compelling explanation for such pricing patterns. For perfect competition, one expects prices to be driven down to cost, while standard nonlinear pricing models suggest the highest-demand consumer will pay the lowest marginal price. Instead, a tendency of consumers to underestimate the variance of their future demand when choosing a tariff provides a more plausible explanation of observed menus of three-part tariffs. Two important biases lead to this tendency: forecasting overconfidence, which has been well documented in the psychology literature, and projection bias, which is described by George Loewenstein, Ted O'Donoghue, and Matthew Rabin (2003).

*MIT Sloan School of Management, 50 Memorial Drive, Cambridge, MA 02142-1347 (e-mail: mgrubb@mit.edu). A previous version of this paper circulated under the title "Screening Overconfident Consumers" (2005). I am very grateful to Jeremy Bulow, Jonathan Levin, and Andrzej Skrzypacz for many valuable discussions of the issues in the paper, and to Katja Seim for help and advice, especially in obtaining data. For helpful comments and suggestions, I would also like to thank three anonymous referees, Susan Athey, Simon Board, Carlos Corona, Liran Einav, Erik Eyster, Bob Gibbons, Richard Holden, Peter Lorentzen, Greg Rosston, Brian Viard, Bob Wilson, and seminar participants at Stanford University, UC-Berkeley, Northwestern University, MIT, Princeton University, Yale University, Columbia University, UCLA, Caltech, Harvard University, and the University of Pennsylvania. I am thankful for financial support from the Taube Scholarship Fund Fellowship through a grant to the Stanford Institute for Economic Policy Research, and from the State Farm Companies Foundation Doctoral Award.

¹ My thanks to an anonymous referee for suggesting this example.

Intuitively, underestimating variance of future demand may lead to tariffs of the form observed because consumers do not take into account the risk inherent in the convexity of the tariffs on the menu. Although the tariffs have a high average cost per unit for consumers who consume far above or far below their allowance, consumers are overly certain that they will choose a tariff with an allowance that closely matches their consumption. Thus, consumers expect to pay a low average price per unit, but sellers profit ex post when consumers make large revisions in either direction.² This intuition is illustrated with a simple example in Section II.

I develop a model of firm pricing when consumers are overconfident. I begin by assuming that consumers are homogenous ex ante, so that there is no screening at the contracting stage and firms offer only a single tariff. In this context, I show that prices will be qualitatively similar to three-part tariffs. Given consumer overconfidence, free disposal, and low marginal costs, consumers will be offered a tariff that involves a range of units offered at zero marginal price, followed by positive marginal prices for additional units. This result holds not only under monopoly, but also under perfect competition. Furthermore, while overconfidence always reduces total surplus, it may increase consumer welfare and reduce monopoly profits. Predicted prices may be fully nonlinear; however, a three-part tariff is always a good approximation when overconfident consumers are primarily uncertain about the volume of desirable units, relative to a fairly consistent value for units that are desirable.

I extend the model to allow for ex ante heterogeneity and screening at the contracting stage via a menu of multiple tariffs. I characterize a monopolist's optimal two-tariff menu given two ex ante types. The qualitative pricing results of the single-tariff model are robust as long as overconfidence is sufficiently high relative to ex ante heterogeneity in average demand. A general characterization of multiple-tariff menus given perfect competition is more difficult. However, specific examples of multiple-tariff menus under perfect competition illustrate the same qualitative tariff features as in the single-tariff model.

I consider three alternative explanations for three-part tariff pricing. First, I consider the flat-rate tariff bias, which encompasses demand overestimation, risk aversion, and the taxi-meter effect (Lambrecht and Skiera 2006). Second, I consider demand underestimation, which is related to quasi-hyperbolic discounting. Third, I consider a monopoly price discrimination explanation which is closely related to Pascal Courty and Hao Li (2000). Although the first two potential alternatives may have important effects on pricing, neither can explain three-part tariff offerings. However, the monopoly price discrimination model does predict three-part tariff pricing given the right type distribution. As the overconfidence and price discrimination models cannot be distinguished based on observed (monopoly) prices, I compare the two explanations using both observed prices and tariff and quantity choices in a particular setting: cellular phone services.

I have obtained billing records for 2,332 student accounts managed by a major US university for a national US cellular phone service provider. The data span 40 of the 41 months February 2002 through June 2005 (December 2002 is missing), and include 32,852 individual bills. I find that customer tariff choices and subsequent usage decisions are not only consistent with the model of overconfidence, but are just what would be expected from overconfident consumers. Moreover, usage patterns suggest that the overconfidence explanation is more appropriate than the price discrimination explanation in this particular application. Specifically, the distribution of usage by customers on a plan with a large number of included minutes strictly first-order stochastically dominates (FOSD) the distribution of usage by customers on a plan with a small number of included minutes. This is inconsistent with the price discrimination model given three-part tariff pricing.

² According to a pricing manager at a top US cellular phone service provider, "people absolutely think they know how much they will use and it's pretty surprising how wrong they are."

The paper proceeds as follows. Section I discusses related literature. Section II illustrates the intuition for the results with a simple example. Section III presents and analyzes the single-tariff model, and Section IV extends the analysis to multiple-tariff menus (focusing on a menu of two tariffs). Potential alternative explanations are described in Section V, and then tested empirically in Section VI. Finally, Section VII concludes. (Supplementary material referred to in the paper is contained in the Web Appendix, available at <http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.5.1770>.)

I. Related Literature

A. *Nonlinear Pricing*

Any model that explains the use of three-part tariffs should capture their primary qualitative feature: included quantities at zero marginal price followed by steep marginal charges. Standard nonlinear pricing models (Michael Mussa and Sherwin Rosen 1978; Eric Maskin and John Riley 1984; Robert B. Wilson 1993) predict marginal cost pricing for the last unit, and higher marginal prices for all lower quantities. They cannot explain marginal charges that are at or below marginal cost for low quantities, but are significantly more expensive at higher quantities.³

While standard screening models are static, reality is dynamic. Several papers explicitly model two-stage screening in which agents choose a contract using a signal about their preferences, but later learn their true preferences before making a quantity choice (David P. Baron and David Besanko 1984; Michael H. Riordan and David E. M. Sappington 1987; Eugenio J. Miravete 1996; Courty and Li 2000; Miravete 2005). Although none of this research specifically addresses three-part tariff pricing, this branch of the nonlinear pricing literature is a natural reference point.

In particular, the single-tariff model presented in Section III is closely related to the standard monopoly screening problem. The important differences in this paper are the incorporation of consumer overconfidence, free disposal, and an ex ante participation constraint, which together predict pricing qualitatively similar to a single three-part tariff. Moreover, both the multiple-tariff monopoly model with overconfidence presented in Section IV and the price discrimination model with common priors discussed in Section V are closely related to Courty and Li (2000). These two models differ from Courty and Li (2000) by incorporating continuous demand with declining marginal valuations and free disposal, and either by adding overconfidence (Section IV) or by considering alternative type distributions (Section V).

I incorporate overconfidence by relaxing the common prior assumption. There are a few related papers that also investigate firm pricing when consumers have biased priors about their future demand. Kfir Eliaz and Ran Spiegler (2008) consider a model where consumers have biased priors but only two types of ex post demand—high or low—in contrast to the continuous demand case studied here. The difference is important because although Eliaz and Spiegler (2008) can capture consumers who over- or underestimate average demand, they cannot capture overconfident consumers who underestimate the likelihood of both upper and lower tails

³ Of course, prices on a particular tariff for quantities that are never chosen may be somewhat arbitrary. In a static screening model, all that matters in a tariff menu is the lower envelope of tariffs on the menu. Segments of tariffs that are above that minimum may be set arbitrarily, for instance, to include regions of zero marginal price. This does not explain the structure of cell phone tariffs, however. First, zero marginal price regions are typically part of the lower envelope of tariffs on the menu. What is more, customer billing data show that usage falls within the zero marginal price regions of tariffs approximately 80 percent of the time, and then on average reaches only half of the included allowance (see Section VI). This undermines the interpretation that a menu of three-part tariffs is an implementation of Mussa and Rosen's (1978) optimal fully nonlinear tariff, and other explanations based on weak optimality (e.g., Walter Y. Oi 1971; Sissel Jensen 2006), since these all predict that every customer consumes at or beyond his or her included allowance.

of demand by overweighting the center of the distribution. Allowing for more than two ex post types is also crucial (along with free disposal and consumer satiation) to generate three-part tariff pricing.⁴ (Eliasz and Spiegler's 2006 and 2008 papers are similar, but the earlier analysis examines dynamically inconsistent consumers.)

Andreas Uthemann (2005) considers a model that, like mine, has continuous ex post demand, but, unlike mine (and similar to Eliasz and Spiegler 2008), does not allow for free disposal or satiation. As a result, the optimal tariffs under both monopoly and competition are qualitatively different from those in this paper, and in particular have strictly positive marginal prices everywhere.⁵ Finally, Alvaro Sandroni and Francesco Squintani (2007) characterize equilibrium pricing in a competitive Michael Rothschild and Joseph E. Stiglitz (1976) insurance market where a fraction of customers underestimate their risk of suffering a loss. A similarity between our models is that, unlike Eliasz and Spiegler (2008) or Uthemann (2005), we allow firms' beliefs about the "true" distribution of demand to vary across consumers, as well as heterogeneity in consumers' own beliefs. However, like Eliasz and Spiegler (2008), Sandroni and Squintani (2007) assume that there are only two consumer types ex post—in this case those who suffer a loss and those who do not.

There are a number of other related papers that consider the effect of alternate consumer biases or nonstandard preferences on optimal nonlinear pricing (Stefano DellaVigna and Ulrike Malmendier 2004; Sharon M. Oster and Fiona M. Scott Morton 2005; Susanna Esteban and Eiichi Miyagawa 2005; Xavier Gabaix and David Laibson 2006; Esteban, Miyagawa, and Matthew Shum 2007). I discuss several of these as potential explanations of observed pricing patterns in Section V.

B. *Overconfidence*

Loewenstein et al. (2003) present a variety of laboratory and survey evidence demonstrating the prevalence of projection bias, and Michael Conlin, O'Donoghue, and Timothy J. Vogelsang (2007) document projection bias in the field using catalog sales and returns data. Individuals who exhibit this bias overestimate the degree to which their future tastes will resemble their current tastes, and therefore tend to underestimate the variance of their future demand. Moreover, a significant body of experimental evidence shows that individuals are overconfident about the precision of their own predictions when making difficult⁶ forecasts (e.g., Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips 1982). In other words, individuals tend to set overly narrow confidence intervals relative to their own confidence levels. A typical psychology study might pose the following question to a group of subjects: "What is the shortest distance between England and Australia?" Subjects would then be asked to give a set of confidence intervals centered on the median. A typical finding is that the true answer lies outside a subject's 98 percent confidence interval about 30 to 40 percent of the time. Ex post tariff-choice "mistakes" made by cellular phone customers are consistent with such overconfidence, as documented in Section VI.

⁴ Technically, one needs at least three types to separate overconfidence from overoptimism. Similarly, one needs many ex post equilibrium quantities to pin down uniquely the shape of the optimal contract.

⁵ On the technical side, an important difference is that Uthemann (2005) characterizes the solution to a relaxed problem without providing sufficient conditions on model primitives to guarantee that it coincides with the solution to the full problem. As I show in Web Appendix B, the two solutions do not coincide when consumer overconfidence is high.

⁶ Predicting one's future demand for minutes is a relatively difficult task, at least for new cellular phone users. Consumers must predict not only the volume of outgoing calls they will make, but also the number of incoming calls they will receive.

II. Illustrative Example

A simple example illustrates my main results. Assume that a supplier has a constant marginal cost of 5 cents per minute and a fixed cost of \$50 per customer.⁷ Consumers value each additional minute of consumption at 45 cents up to some satiation point, beyond which they value further minutes at 0 cents. When consumers sign up for a tariff in period 1, they are homogeneously uncertain about their satiation points. Then in period 2, consumers learn their satiation points and use this information to make their consumption choices. In particular, assume that one-third of consumers learn that they will be satiated after 100 minutes, one-third after 400 minutes, and the remaining third after 700 minutes.

If consumers and the supplier share this prior belief, then it is optimal for the firm to charge a marginal price equal to the marginal cost of 5 cents per minute.⁸ Under monopoly, the firm extracts all the surplus via a fixed fee of \$160, earning profits of \$110 per customer. Under perfect competition, the firm charges a fixed fee of \$50, leaving \$110 in surplus to consumers.

If consumers are overconfident, however, marginal cost pricing is no longer optimal. For instance, if all consumers are extremely overconfident and believe that they will be satiated after 400 minutes with probability 1, then it is optimal to charge 0 cents per minute for the first 400 minutes, and 45 cents per minute thereafter. In other words, it is optimal to have 400 “included” minutes in the tariff.

Under monopoly, the firm charges a fixed fee of \$180, earning expected profits of \$155 per customer. Ex ante consumers expect to receive zero surplus, but on average ex post realize a loss of \$45. Under perfect competition, the firm charges a fixed fee of \$25, and consumers expect to receive \$155 in surplus, but actually realize only \$110. Consumer overconfidence allows the creation ex ante of an additional \$45 in perceived consumer surplus, which is never realized ex post.

To see why this tariff is optimal, consider the pricing of minutes 100–400 and 400–700 separately. On the one hand, overconfident consumers believe that they will consume minutes 100–400 with probability 1, while the firm knows that they will actually consume them only with probability $\frac{2}{3}$. As a result, reducing the marginal price of minutes 100–400 from 5 cents to 0 cents is perceived differently by the firm and consumer. The consumer views this as a \$15 price cut and will be indifferent if the fixed fee is increased by \$15. The firm, however, recognizes this as only a \$10 revenue loss, and will be better off by \$5 if the fixed fee is raised by \$15.

On the other hand, overconfident consumers believe they will consume minutes 400–700 with probability 0, while the firm knows that they will actually consume them with probability $\frac{1}{3}$. Therefore from the consumer’s perspective, increasing the marginal price of minutes 400–700 from 5 cents to 45 cents does not affect the expected price paid. The firm, however, views this as an increase in expected revenues of \$40.

Essentially, the firm finds it optimal to sell the first 400 minutes up front to overconfident consumers. Then, in the second period, the firm buys back minutes 100–300 from the low demand consumers at the monopsony price of 0 cents per minute, and sells minutes 400–700 to high demand consumers at the monopoly price of 45 cents per minute.

Note that, in this example, a monopolist earns higher profits from overconfident consumers, making them worse off than consumers with correct priors. Under competition, however,

⁷ Fixed costs per customer may arise due to billing costs, a subsidy for a new phone, or customer acquisition fees paid to retailers.

⁸ Note that this is only one of a continuum of optimal pricing structures which all implement the efficient allocation. Were demand curves not rectangular and were there a continuum of types, then marginal cost pricing would be uniquely optimal.

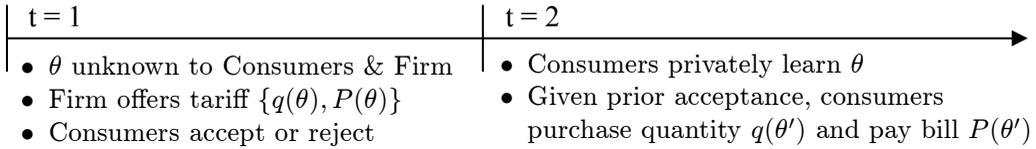


FIGURE 1. TIMELINE

overconfident consumers are equally as well off as consumers with correct priors. Neither result is true in general; rather, both follow from the specific form of preferences assumed (see Section IIIH).

III. Single-Tariff Analysis

A. Model

Game players are a firm, or multiple firms in the case of perfect competition, and a continuum of consumers. Timing of the game (Figure 1) differs from a standard screening model. At the contracting stage ($t = 1$) consumers are homogeneous and do not know their future demand type θ . The firm offers tariff $\{q(\theta), P(\theta)\}$, which describes a purchase quantity and payment pair intended for each type θ . Consumers accept or reject based on their prior belief over θ at $t = 1$. Finally, consumers privately learn θ and choose to purchase quantity $q(\theta')$ at $t = 2$.

The base assumptions about production and preferences match those of a standard screening model. A firm's profits Π are given by revenues P less production costs $C(q)$, which are increasing and convex in quantity delivered, q . Consumers' utility U is equal to their value of consuming q^c units, $V(q^c, \theta)$, less their payment to the firm, P . Consumers' marginal value of consumption V_q is strictly decreasing in consumption q^c , and strictly increasing in consumers' type θ . The outside option of all consumers is the same and normalized to zero: $V(0, \theta) = 0$.

I make an additional assumption concerning consumer preferences, which would not be relevant in a standard model: consumers have a finite satiation point, $q^S(\theta) \equiv \arg \max_{q^c \geq 0} V(q^c, \theta)$, beyond which they may freely dispose of unwanted units. Hence, consumer type θ who purchases $q(\theta')$ units will consume only the minimum of $q(\theta')$ and $q^S(\theta)$, and will receive consumption value $V(\min\{q^S(\theta), q(\theta')\}, \theta)$.⁹

The key assumption of the model, which deviates sharply from a standard model, is that consumers underestimate the variance of their future demand θ . This is either because they are overconfident about the accuracy of their forecasts of θ , or because they are subject to projection bias. Thus, while the firm knows¹⁰ that consumer demand θ follows cumulative distribution $F(\theta)$, consumers have the prior belief that θ follows $F^*(\theta)$. Moreover, the firm knows that consumers

⁹ The satiation point $q^S(\theta)$ is the point at which type θ 's (strictly decreasing) marginal value of consumption becomes negative. Finite satiation implies that consumers need not purchase all units with zero marginal price. Rather than explicitly allowing for free disposal, it would have been equivalent to assume directly that consumers have value function $\bar{V}(q, \theta) = V(\min\{q, q^S(\theta)\}, \theta)$ for which the marginal value of consumption is zero beyond the finite point $q^S(\theta)$.

¹⁰ Strictly speaking, there is no need to assume that either the firm's prior or the consumer's prior is correct, except in order to make statements about welfare. The interpretation maintained throughout this paper is that the firm's beliefs are correct and the consumers' beliefs are incorrect. A larger game is imagined in which the firm quickly learns the true distribution of types of new consumers by observation of its large number of existing customers. New consumers, however, are overconfident and believe they know more about their own type than they really do, as described in (A*).

are overconfident, so will take this into account when designing its tariff offering. Finally, the disagreement between the firm and consumers is captured by assumption A*:

ASSUMPTION A*:¹¹ $F^*(\theta)$ crosses $F(\theta)$ once from below at θ^* .

An interesting special case of A* is where consumers and the firm agree on the mean of θ , in which case $F(\theta)$ is a mean preserving spread of $F^*(\theta)$ and consumers underestimate the variance of their future demand.

Within the context of this model, the equilibrium tariff, or allocation and payment pair $\{\hat{q}(\theta), \hat{P}(\theta)\}$, will be characterized under both monopoly and perfect competition. This analysis requires several more technical assumptions. As is standard, it is assumed that $V(q, \theta)$ is thrice continuously differentiable, $C(q)$ and $F(\theta)$ are twice continuously differentiable, $F^*(\theta)$ is continuous and piecewise smooth, consumption is nonnegative, and total surplus is initially strictly increasing in q . The firm's prior $F(\theta)$ has full support over $[\underline{\theta}, \bar{\theta}]$, a range that includes the support of consumers' prior $F^*(\theta)$.

Finally, I restrict the partial derivative $V_{qq\theta}$ by equation (1), where $q^{FB}(\theta) \equiv \text{argmax}_q [V(q, \theta) - C(q)]$ denotes the first-best allocation. This ensures that the virtual surplus function described in Proposition 1 is strictly quasi-concave. Note that when marginal costs are zero (the focus of the paper), consumer satiation coincides with the first-best allocation, and equation (1) reduces to the standard assumption: $V_{qq\theta} \geq 0$ (e.g., see Drew Fudenberg and Jean Tirole 1991, ch. 7).

$$(1) \quad V_{qq\theta}(q, \theta) \begin{cases} \geq 0, & 0 < q < q^{FB}(\theta) \\ \leq 0, & q^{FB}(\theta) < q < q^S(\theta) \end{cases} .$$

Parallel demand curves ($V_{q\theta} = 0$) are a special case of equation (1) for which it is without further loss of generality to set $V_{qq\theta} = 0$ by appropriate normalization of θ . The consumers' value function may then be written as $V(q^c, \theta) = v(q^c) + q^c \theta$. Moreover, consumers who correctly predict the mean of θ also correctly predict their mean value of each unit. Another special case of equation (1) are demand curves that are translated horizontally by θ and are concave in quantity if marginal costs are zero, or are concave below the first-best allocation and convex above if constant marginal costs are strictly positive. Unlike parallel demand curves, these can capture

¹¹Note that Assumption A* corresponds closely to the two documented biases, forecasting overconfidence and projection bias, from which it is motivated. For instance, the special case of Assumption A* , where $F^*(\theta)$ is given by the equation below for some $\alpha \in (0, 1)$, exactly matches Loewenstein et al.'s (2003) formalization of projection bias:

$$F^*(\theta) = \begin{cases} (1 - \alpha), F(\theta) & \theta < \theta^* \\ (1 - \alpha), F(\theta) + \alpha & \theta \geq \theta^* \end{cases} .$$

In this case, θ^* would be interpreted as a consumer's current taste for consumption when making his or her participation decision at $t = 1$. (This is not how Loewenstein et al. (2003) present their model, but it is straightforward to show the equivalence, as they hint in their Footnote 8.)

Further, Assumption A* guarantees that any confidence interval drawn by an individual that includes θ^* will be overly narrow. Conversely, if all of an individual's perceived confidence intervals that include θ^* are strict subsets of the true confidence intervals, Assumption A* must hold. If we think of θ^* as a central point such as the median, this provides a strong link to the studies of forecasting overconfidence.

Finally Assumption A* is closely related to second-order stochastic dominance (SOSD). Let μ^* and μ be the mean values of distributions F^* and F , respectively. Given A*, F^* SOSD F if and only if $\mu^* \geq \mu$ (Giora Hanoach and Haim Levy 1969, Theorem 3).

preferences close to $V(q^c, \theta) = p \min \{q^c, \theta\}$, which describe consumers who have a relatively consistent and certain value for desirable units, but are uncertain about the quantity desired.

B. Defining the Problem

Invoking the standard revelation principle, the equilibrium monopoly tariff $\{q^M(\theta), P^M(\theta)\}$ must solve the following constrained profit maximization problem:

$$\max_{\substack{P(\theta) \\ q(\theta) \geq 0}} E[P(\theta) - C(q(\theta))]$$

such that:

1. Global Incentive Compatibility $U(\theta, \theta) \geq U(\theta, \theta') \quad \forall \theta, \theta' \in [\underline{\theta}, \bar{\theta}]$,
2. Consumer Participation¹² $E^*[U(\theta)] \geq 0$.

The monopolist’s problem is similar to that in a standard screening model. The monopolist’s objective is the same: to maximize expected profits. Moreover, at $t = 2$ when consumers privately learn their types, it must be optimal for consumers to truthfully reveal their types by self-selecting appropriate quantity—payment pairs from the tariff. Thus, the standard incentive compatibility constraint applies: the utility $U(\theta, \theta') \equiv V(\min \{q^S(\theta), q(\theta')\}, \theta) - P(\theta')$ of a consumer of type θ who reports θ' at $t = 2$ must be weakly below the utility $U(\theta) \equiv U(\theta, \theta)$ of a consumer of type θ who reports truthfully at $t = 2$.

There are two important deviations from a standard screening model. First, free disposal is explicitly incorporated through consumer preferences, which depend on the consumed quantity $\min \{q^S(\theta), q(\theta')\}$ rather than the purchased quantity $q(\theta')$. Second, contracting occurs ex ante, at which time consumers’ beliefs differ from those of the firm. Thus the ex ante participation constraint requires that consumers’ perceived expected utility $E^*[U(\theta)]$ must be positive, but puts no constraint on their true expected utility $E[U(\theta)]$. The difference in priors between consumers and the firm creates a wedge separating the expected utility consumers believe they are receiving from the expected utility the firm believes it is actually providing.

Invoking the revelation principle a second time, the equilibrium tariff $\{q^C(\theta), P^C(\theta)\}$ under perfect competition must solve the following closely related constrained maximization problem:

$$\max_{\substack{P(\theta) \\ q(\theta) \geq 0}} E^*[U(\theta)]$$

such that:

1. Global Incentive Compatibility $U(\theta, \theta) \geq U(\theta, \theta') \quad \forall \theta, \theta' \in [\underline{\theta}, \bar{\theta}]$,
2. Producer Participation $E[P(\theta) - C(q(\theta))] \geq 0$.

As under monopoly, the equilibrium tariff must satisfy incentive compatibility constraints. The difference is that the objective function and participation constraints are reversed. Under perfect competition the equilibrium tariff maximizes consumers’ perceived expected utility

¹² Expectations taken with respect to the consumers’ prior $F^*(\theta)$ are denoted by a superscript * on the expectations operator.

subject to firm participation, as otherwise there would be an opportunity for profitable entry. In contrast, under monopoly, firm payoff is maximized subject to consumer participation.

C. Simplifying the Problem

The initial step in simplifying the problem is to recognize that there will never be any reason for firms to induce a consumer to purchase beyond her satiation point, as she would simply dispose of unwanted additional units. By initially selling the consumer her satiation quantity at the same price, the consumer would have been equally well off, incentives constraints of other consumers would have been weakly relaxed, and the firm could have reduced production costs. This result is stated formally in Lemma 1.

LEMMA 1: *If the pair $\{\hat{q}(\theta), \hat{P}(\theta)\}$ is an optimal tariff under either monopoly or perfect competition, then the pair $\{\min\{\hat{q}(\theta), q^S(\theta)\}, \hat{P}(\theta)\}$ is also optimal. Moreover, if production costs are strictly increasing, then $\hat{q}(\theta) \leq q^S(\theta)$ almost everywhere.*

PROOF:

Follows from the assumption of free disposal. See Appendix.

I will focus on equilibria in which firms offer allocations no higher than consumers' satiation points. Given Lemma 1, this is without loss of generality when marginal costs are strictly positive. When marginal costs are zero, this refinement simply selects the limiting equilibrium as marginal costs approach zero. As a result, rather than separately tracking equilibrium purchases $q(\theta)$ and consumption $\min\{q(\theta), q^S(\theta)\}$, knowing that they will be the same I can work with a single quantity $q(\theta)$ by imposing a satiation constraint $q(\theta) \leq q^S(\theta)$.

Having reduced equilibrium purchase and consumption quantities to a single function $q(\theta)$, the problem can be further simplified following the standard approach. The first step, introduced by James A. Mirrlees (1971), is to replace the global incentive compatibility constraint with the joint constraints of local incentive compatibility and monotonicity. The second step is to recognize that under either monopoly or perfect competition, the relevant participation constraint must bind. Now, for every allocation $q(\theta)$ there is a unique payment function $P(\theta)$ which satisfies local incentive compatibility and meets the relevant participation constraint with equality.¹³ Both monopoly and perfect competition problems may then be simplified by substituting these two constraints in place of payments $P(\theta)$ in the objective function.

Completing the described substitution for both monopoly and perfect competition reveals a beautiful aspect of the two problems. The transformed objective functions, now expressed solely as a function of the allocation $q(\theta)$, are identical under both monopoly and perfect competition. (This is true because consumers are homogeneous at the time of contracting.) In particular, in both market scenarios $q(\theta)$ maximizes an expected virtual surplus $E[\Psi(q(\theta), \theta)]$, subject to the remaining nonnegativity, monotonicity, and satiation constraints. Expected virtual surplus is equal to the sum of expected true surplus, $S(q(\theta), \theta) \equiv V(q(\theta), \theta) - C(q(\theta))$, and a "fictional surplus," which is the difference between the expected utility $E^*[U(\theta)]$ consumers believe they are receiving and the expected utility $E[U(\theta)]$ the firm believes it is delivering (equation (2)).

¹³ This payment function can be found first by expressing payments in terms of consumer utility: $P(\theta) = V(\min\{q(\theta), q^S(\theta)\}, \theta) - U(\theta)$. Next, local incentive compatibility requires that $U(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} V_{\theta}(q^c(z), z) dz$, which pins down payments up to a constant $U(\underline{\theta})$. Finally, binding participation constraints determine the constant $U(\underline{\theta})$.

Moreover, having substituted local incentive compatibility and participation constraints in place of payments, fictional surplus is given by equation (3):

$$(2) \quad E[\Psi(q(\theta), \theta)] = E[S(q(\theta), \theta)] + E^*[U(\theta)] - E[U(\theta)],$$

$$(3) \quad E^*[U(\theta)] - E[U(\theta)] = E\left[V_\theta(q(\theta), \theta) \frac{F(\theta) - F^*(\theta)}{f(\theta)}\right].$$

When consumers and the firm share the same prior ($F^*(\theta) = F(\theta)$), fictional surplus is zero, so the equilibrium tariff maximizes expected surplus $E[S(q(\theta), \theta)]$. This implies first-best allocation and marginal payment equal to marginal cost. When consumers are overconfident, however, fictional surplus need not be zero, and may distort the equilibrium allocation away from first best, and marginal pricing away from marginal cost. These distortions, and thus the equilibrium allocation $\hat{q}(\theta) = q^M(\theta) = q^C(\theta)$, will be identical under monopoly and perfect competition. As a result, marginal pricing, which is pinned down jointly by the allocation and local incentive compatibility, will be the same across market conditions. The only variation in pricing will be a higher fixed fee under monopoly, due to the difference in participation constraints across market conditions.¹⁴ Proposition 1 summarizes these results precisely.

PROPOSITION 1: *Under both monopoly and perfect competition:*

(i) *Equilibrium allocations are identical and maximize expected virtual surplus:*

$$\hat{q}(\theta) = \arg \max_{\substack{q(\theta) \in [0, q^s(\theta)] \\ q(\theta) \text{ non-decreasing}}} E[\Psi(q(\theta), \theta)].$$

$$(4) \quad \Psi(q, \theta) \equiv V(q, \theta) - C(q) + V_\theta(q, \theta) \frac{F(\theta) - F^*(\theta)}{f(\theta)}.$$

(ii) *Payments differ only by a fixed fee and are given by:*

$$(5) \quad P^M(\theta) = V(\hat{q}(\theta), \theta) - \int_{\underline{\theta}}^{\theta} V_\theta(\hat{q}(z), z) dz + E\left[V_\theta(\hat{q}(\theta), \theta) \frac{1 - F^*(\theta)}{f(\theta)}\right].$$

$$(6) \quad P^C(\theta) = P^M(\theta) - E[\Psi(\hat{q}(\theta), \theta)].$$

¹⁴ The main results are easily extended to imperfect competition in which firms are differentiated by location and consumers' transportation costs d are independent of consumption or type θ (for example, $V(q, \theta, d) = V(q, \theta) - d$). Equilibrium allocations and marginal prices would be identical to those in the current model, which maximize expected virtual surplus. Firms would compete with each other through the fixed fees, which would drop with the level of competition. In contrast, distortions of price away from marginal cost in a standard price discrimination model disappear with increasing competition (Lars A. Stole 1995). The extension is straightforward as, apart from their brand preferences, consumers are homogeneous at the time of contracting.

PROOF:

Outlined in the text above. For a more formal restatement of the proposition, see the Appendix. For a more detailed proof, see Grubb (2008).

D. Equilibrium Allocation

Further characterization of the equilibrium allocation follows the standard approach. First, the solution $q^R(\theta)$ to a relaxed problem (equation (7)) that ignores the monotonicity constraint is characterized. Second, any nonmonotonicities in $q^R(\theta)$ are “ironed out.” Implications about pricing can then be drawn based on part (ii) of Proposition 1.

$$(7) \quad q^R(\theta) \equiv \arg \max_{q(\theta) \in [0, q^s(\theta)]} \Psi(q, \theta).$$

LEMMA 2:

- (i) *The relaxed solution $q^R(\theta)$ is a continuous function. It is characterized by the first-order condition $\Psi_q(q, \theta) = 0$ except where satiation or nonnegativity constraints bind.*
- (ii) *The equilibrium allocation $\hat{q}(\theta)$ is continuous. On any interval over which the monotonicity constraint is not binding, the equilibrium allocation is equal to the relaxed allocation: $\hat{q}(\theta) = q^R(\theta)$.*

PROOF:

Part (i) is true because equation (1) ensures Ψ is strictly quasi-concave in q . The proof of part (ii) is omitted as it closely follows ironing results for the standard screening model (see Appendix).

Lemma 2 closely parallels analogous results in standard screening models. The important point is that the equilibrium allocation $\hat{q}(\theta)$ is continuous and equal to the relaxed allocation $q^R(\theta)$ where the monotonicity constraint is not binding. This fact is useful because it implies that the relaxed solution $q^R(\theta)$ determines marginal prices (Proposition 2).

When overconfidence is limited, and consumers’ beliefs are close to those of the firm, the relaxed solution will be strictly increasing. However, when consumers are extremely overconfident, the relaxed solution will violate the monotonicity constraint (Web Appendix B, Proposition 6). This is because the relaxed solution is distorted upward (weakly above first best) below θ^* , but downward (below first best) above θ^* (Web Appendix B, Proposition 5). When overconfidence is sufficiently high, the distortions in opposing directions are large enough on either side of θ^* that the relaxed solution must be strictly decreasing at θ^* . Thus, to avoid excluding interesting cases, Web Appendix B completes the characterization of ironed allocations (Lemma 4).

E. Pricing Implications

Together, Proposition 1 and Lemma 2 characterize the equilibrium allocation $\hat{q}(\theta)$ consumed and payment $\hat{P}(\theta)$ paid by each type θ . Rather than using a direct revelation mechanism, however, in practice this tariff will be implemented by setting price as a function of quantity. Let $\hat{\theta}(q) \equiv \inf \{ \theta : \hat{q}(\theta) = q \}$. Then (with a slight abuse of notation) the tariff will be implemented by $\hat{P}(q) = \hat{P}(\hat{\theta}(q))$. It is now possible to draw implications about pricing using Proposition 1.

PROPOSITION 2: *The equilibrium payment $\hat{P}(q)$ is a continuous and piece-wise smooth function of quantity. There may be kinks in the payment function where marginal price increases discontinuously. These kinks occur where the monotonicity constraint binds and an interval of types “pool” at the same quantity. For quantities at which there is no pooling, because the monotonicity constraint does not bind, marginal price is given by*

$$(8) \quad \frac{d\hat{P}(q)}{dq} = V_q(q, \hat{\theta}(q)) = \max \left\{ 0, C_q(q) + V_{q\theta}(q, \hat{\theta}(q)) \frac{F^*(\hat{\theta}(q)) - F(\hat{\theta}(q))}{f(\hat{\theta}(q))} \right\}.$$

PROOF:

See Appendix.

As it is assumed that $V_{q\theta}$ is strictly positive and $f(\theta)$ is finite, Proposition 2 allows marginal price to be compared to marginal cost based on the sign of $[F^*(\theta) - F(\theta)]$. In particular, the sign of $[\hat{P}_q(q) - C_q(q)]$ is equal to the sign of $[F^*(\theta) - F(\theta)]$ except when $F^*(\theta) < F(\theta)$ and marginal cost is zero, as then marginal price is also zero. This is informative about equilibrium pricing, as Assumption A* dictates the sign of $[F^*(\theta) - F(\theta)]$ above and below θ^* .

Define \underline{q} , Q , and \bar{q} to be the equilibrium allocations of types $\underline{\theta}$, θ^* , and $\bar{\theta}$, respectively:

$$\{\underline{q}, Q, \bar{q}\} \equiv \{\hat{q}(\underline{\theta}), \hat{q}(\theta^*), \hat{q}(\bar{\theta})\}.$$

Relevant implications of Proposition 2 are then summarized in Corollary 1.

COROLLARY 1: *Given A*, for quantities at which there is no pooling:*

(i) *If marginal cost is zero for all q , then*

$$\hat{P}_q(q) = 0, \quad q \in (\underline{q}, Q) \cup \{\underline{q}, Q, \bar{q}\},$$

$$\hat{P}_q(q) > 0, \quad q \in (Q, \bar{q}).$$

(ii) *If marginal cost is strictly positive for all q , then*

$$\hat{P}_q(q) = C_q(q) > 0, \quad q \in \{\underline{q}, Q, \bar{q}\},$$

$$C_q(q) > \hat{P}_q(q) \geq 0, \quad q \in (\underline{q}, Q),$$

$$\hat{P}_q(q) > C_q(q) > 0, \quad q \in (Q, \bar{q}).$$

PROOF:

Follows directly from Proposition 2, Assumption A*, and $\hat{q}(\theta)$ nondecreasing.

Corollary 1 shows that when marginal costs are zero, marginal price will be zero below some included allowance Q , and positive thereafter. When marginal costs are strictly positive, marginal price will initially be positive, but will fall below marginal cost and may be zero for some early range of consumption. The following section illustrates these results with numerical examples.

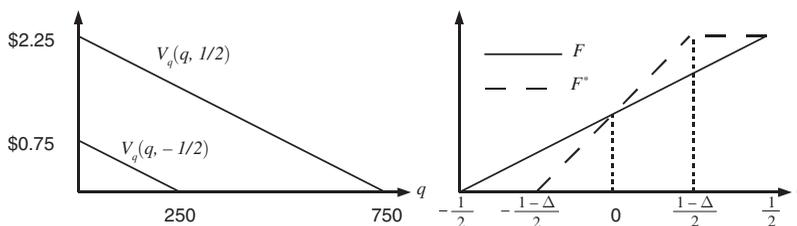


FIGURE 2. INVERSE DEMAND CURVES AND PRIORS IN EXAMPLE 1

It is reasonable to assume that the marginal cost of providing an extra minute of call time to a cellular phone customer is small. Therefore, given overconfident consumers, the equilibrium tariff bears a striking qualitative resemblance to those offered by cellular phone service providers. Both predicted equilibrium tariffs and observed tariffs involve zero marginal price up to some included minute limit Q and become positive thereafter. (In contrast to observed tariffs, the model predicts that marginal price is always equal to marginal cost for the last unit sold. However, this result is not robust to ex ante heterogeneity. See Section IV and Figure 6.)

The intuition for the result is as follows. If consumers are overly confident that their future consumption will be near Q minutes, they will overestimate the probability of consuming initial units. Thus, a firm can overcharge consumers for the first Q minutes through a fixed fee ex ante. By setting a zero marginal price, the firm avoids paying a refund to those consumers who are later surprised by a low level of demand below Q . Overconfident consumers also underestimate the probability of extremely high usage. Thus, a firm cannot extract consumers' true value for high consumption through a fixed fee ex ante. Instead, the firm must wait until consumers learn their true values and charge a marginal fee for high usage above Q . The drawback of such marginal fees is that they inefficiently distort consumption below first best, which limits their optimal size.

Absent free disposal, the firm could do better by setting negative marginal prices on initial units in order to charge consumers for purchasing less than anticipated. However, free disposal ensures that such negative marginal pricing is futile at best, and strictly suboptimal when marginal costs are positive. Consumers can costlessly avoid underusage penalties by purchasing negative marginal price units and disposing of them. It is free disposal that makes zero a sharp lower bound for marginal price in equation (8). (Web Appendix A provides additional intuition based on option pricing.)

F. Example

The implications of Proposition 2 that are summarized in Corollary 1 are best illustrated with figures from specific examples.

EXAMPLE 1: Firms have a fixed cost of \$25 and a constant marginal cost of $c \geq 0$ per unit. Consumers' inverse demand function is linear in q and θ (Figure 2): $V(q, \theta) = \frac{3}{2}q - \frac{3}{2000}q^2 + \frac{3}{2}q\theta$. The firm and consumers' priors are uniform, centered on 0 (Figure 2): F is $U[-\frac{1}{2}, \frac{1}{2}]$ and F^* is $U[-\frac{(1-\Delta)}{2}, \frac{(1-\Delta)}{2}]$. Consumers and the firm both agree that the mean of θ is equal to 0. The parameter $\Delta \in [0, 1]$ measures the percent by which consumers underestimate the standard deviation of θ . For $\Delta = 0$, consumers are not overconfident at all, and share the firm's prior. For $\Delta = 1$, consumers are extremely overconfident and believe $\theta = 0$ with probability one.

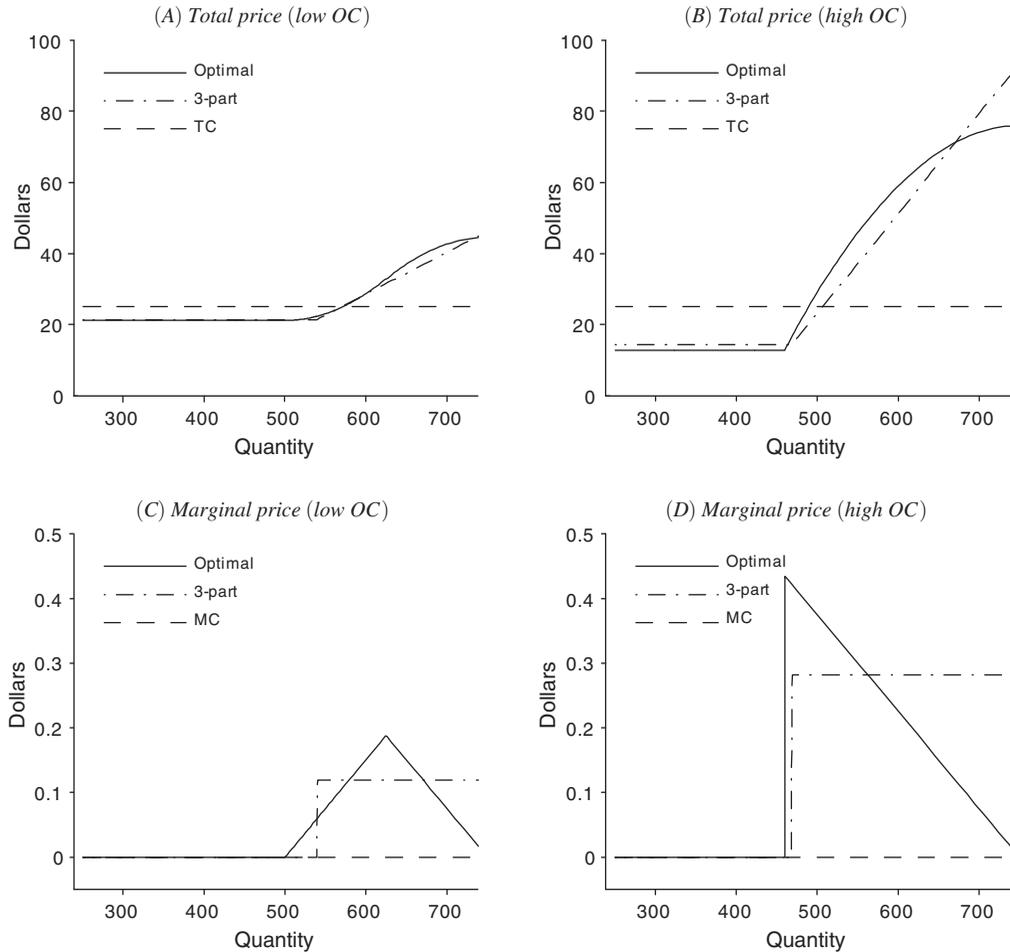


FIGURE 3

Notes: The fully nonlinear equilibrium price and its closest three-part tariff approximation are depicted under perfect competition and zero marginal cost. Prices are shown for low overconfidence ($\Delta = 0.25$) in the left-hand column and for high overconfidence ($\Delta = 0.75$) in the right-hand column. Total prices and costs are shown in the top row, and marginal prices and costs are shown in the bottom row.

Figure 3 illustrates Corollary 1 given zero marginal costs, using the example described above. In the top row, panels A and B show total prices and total costs under perfect competition. There are three curves: optimal total price (solid line), total cost (dashed line), and the best three-part tariff approximation to optimal pricing (dot-dashed line). In the bottom row, panels C and D show marginal prices and marginal costs, under either perfect competition or monopoly. Both the optimal marginal price and its best three-part tariff approximation are shown. In the left-hand column, panels A and C assume low overconfidence, $\Delta = 0.25$, for which there is no pooling. In the right-hand column, panels B and D assume high overconfidence, $\Delta = 0.75$, for which there is pooling at Q .

Figure 3 shows that total payment is constant and marginal price is zero up to some quantity Q . Beyond Q , marginal price is positive. In short, optimal pricing shares important qualitative

TABLE 1—TWO- AND THREE-PART TARIFF APPROXIMATION LOSSES

Marginal cost	\$0	\$0	\$0.035	\$0.035
Overconfidence	low	high	low	high
Two-part loss	\$0.98	\$7.89	\$1.43	\$9.38
Three-part loss	\$0.13	\$1.30	\$0.16	\$1.26
1 – (3-part loss/2-part loss)	87 percent	84 percent	89 percent	87 percent

features of observed three-part tariffs. However the fit is not perfect, because optimal pricing is fully nonlinear. When there is no pooling at Q , total payment increases smoothly beyond Q . When there is pooling at Q , however, the total payment has a kink at Q where marginal price jumps upward discretely. In both cases marginal price falls to zero at the highest quantity \bar{q} .

For comparison, the dot-dashed lines show the best three-part tariff approximations to optimal pricing. In fact, the plots also show the best two-part tariff approximations to optimal pricing since these coincide with the dashed cost curves.¹⁵ Optimal pricing yields a \$0.98 improvement over the best two-part tariff given low overconfidence, and a \$7.89 improvement given high overconfidence. In both cases, the best three-part tariff captures more than 80 percent of this gain, leaving only a \$0.13 or \$1.30 approximation loss for low-overconfidence and high-overconfidence cases, respectively (Table 1). The fact that deviating from marginal cost pricing becomes increasingly valuable with overconfidence is expected since, absent overconfidence, marginal cost pricing is optimal.

Figure 4 shows the same plots given in Figure 3, except that equilibrium payments are plotted for strictly positive marginal cost $c = \$0.035$ rather than zero marginal cost. The plots are similar to those in Figure 3 for quantities above Q . However, as marginal cost is strictly positive, marginal price is strictly positive near \bar{q} . The best three-part tariff approximations offer similarly small approximation losses relative to the best two-part tariffs (Table 1). In the example shown, the satiation constraint does bind and marginal price is zero over some subset of the interval $[q, Q]$. However, were marginal cost higher, the satiation constraint might never bind, and marginal price could be strictly positive at all quantities.

G. Pricing in Practice

As Figures 3 and 4 illustrate, the optimal price characterized by Proposition 2 is fully nonlinear: marginal price may vary for each increment sold beyond the included allowance. In contrast, overage rates are constant for observed tariffs. The fact that observed tariffs are simpler than those predicted by a theory that ignores costs of complexity is not surprising. As Robert B.

¹⁵ Given linear demand curves and correct estimation of the mean ($E^*[\theta] = E[\theta]$), the optimal two-part tariff marginal price is simply marginal cost. To characterize the best two-part tariff approximation in general, let $q^T(\theta, p) = \{q : V_q(q, \theta) = p\}$ be the quantity chosen by type θ given a two-part tariff marginal price p . Then the first-order condition for the optimal p is: $\int_{\theta}^{\bar{\theta}} ([\Psi_q(q^T(\theta, p), \theta)]/[V_{qq}(q^T(\theta, p), \theta)]) f(\theta) d\theta = 0$. To characterize the best three-part tariff approximation, let $\theta(Q, p) = \{\theta : V_q(Q, \theta) = p\}$ be the type who purchases exactly Q units given a marginal price p . The first-order conditions for optimal included allowance Q and overage rate p are $\int_{\theta(Q, 0)}^{\theta(Q, p)} \Psi_q(Q, \theta) f(\theta) d\theta = \int_{\theta(Q, p)}^{\bar{\theta}} ([\Psi_q(q^T(\theta, p), \theta)]/[V_{qq}(q^T(\theta, p), \theta)]) f(\theta) d\theta = 0$. (The optimal initial marginal price is zero in these examples.)

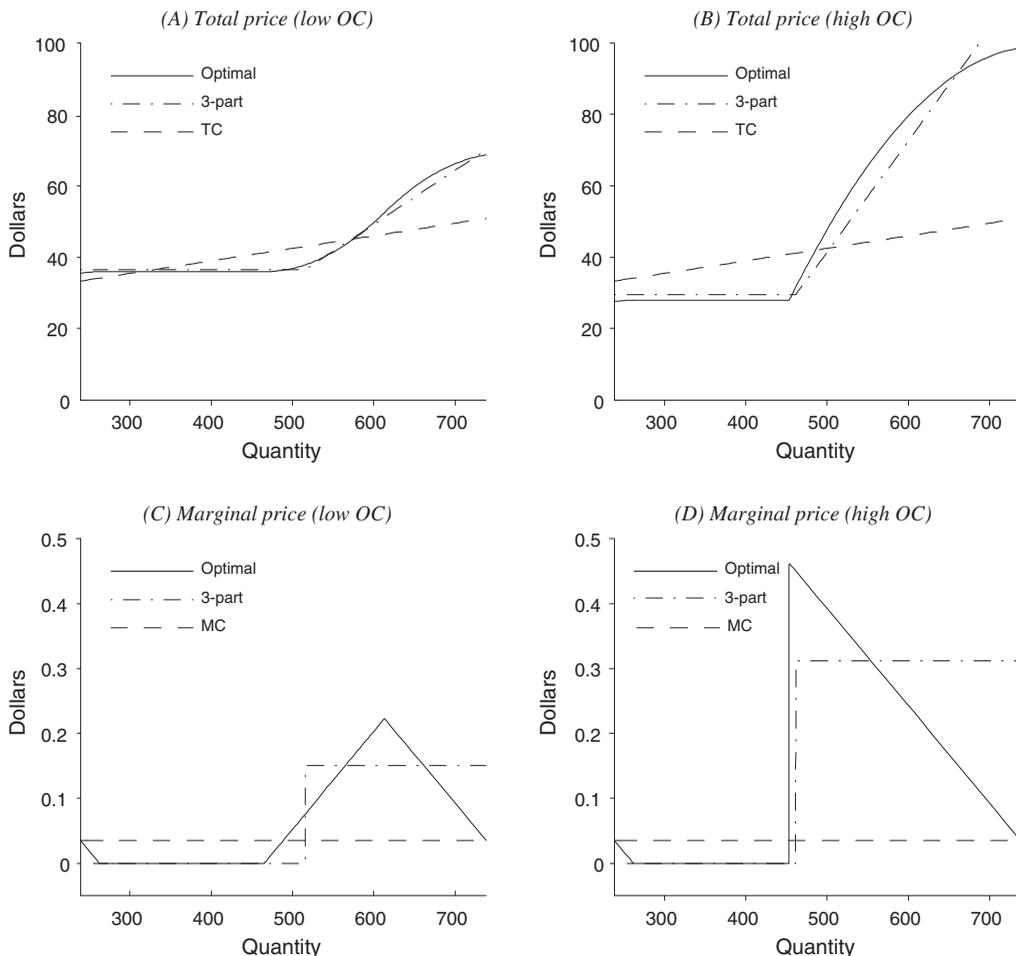


FIGURE 4

Notes: The fully nonlinear equilibrium price and its closest three-part tariff approximation are depicted under perfect competition and positive marginal cost: $c = \$0.035$. Prices are shown for low overconfidence ($\Delta = 0.25$) in the left-hand column and for high overconfidence ($\Delta = 0.75$) in the right-hand column. Total prices and costs are shown in the top row, and marginal prices and costs are shown in the bottom row.

Wilson (1993) points out, “In practice, ... a different price for each increment is differentiation too fine to justify the transaction costs incurred by customers and the firm.”¹⁶

¹⁶ Costs arise for at least two reasons. First, a tariff with many marginal prices would be costly to communicate to consumers, costly for consumers to evaluate, and costly to implement in firm billing systems. Second, the cost of acquiring enough information about demand to distinguish optimal prices for one unit from the next could be substantial. Most firms that charge nonlinear prices specify n -part tariffs, which consist of a fixed fee and $(n - 1)$ intervals of constant marginal price. (The three-part tariffs I focus on are a special case in which the first interval of constant marginal price sets marginal price equal to zero.) Wilson (1993) shows that “The profit and total surplus foregone by using an optimal n -part tariff rather than an optimal tariff with continuously varying prices is of order $1/n^2$ for large values of n ; thus, cost considerations need not be large for a tariff with only a few parts to be optimal.” Given his assumption that monotonicity is not binding, Wilson’s (1993) asymptotic result extends to this setting.

Whether a three-part tariff is an optimal approximation depends on complexity costs, consumer preferences, and beliefs. I now show, however, that a three-part tariff is an excellent approximation whenever overconfident consumers are primarily uncertain about the quantity of desirable units, and tend to have a consistent and certain value for units that are desirable. Moreover, a generalization of Example 1 in Section III F suggests that three-part tariffs may provide large improvements over two-part tariffs relative to the remaining approximation loss for a broader range of preferences.

Consider a generalization of the illustrative example in Section II. If overconfident consumers value the first θ units at p each and are satiated thereafter, $(V(q, \theta) = p \min \{q, \theta\})$, then the optimal contract is a three-part tariff that includes θ^* units at zero marginal price followed by a constant overage rate p . A constant overage rate is optimal regardless of specific beliefs of firms and consumers because although consumers are uncertain *ex ante* about how much they will want to consume, they have a uniform and certain valuation p for any units they do eventually find desirable. Proposition 3 shows that this is not a knife-edge result: if overconfident consumers have preferences close to $V(q, \theta) = p \min \{q, \theta\}$ (so that they are primarily uncertain about their purchase volume, rather than their value for future purchases), then three-part tariffs offer large improvements over two-part tariffs, and closely approximate optimal pricing.

PROPOSITION 3: *Assume that marginal cost is zero, and that consumers' preference $V(q, \theta)$ is close to $p \min \{q, \theta\}$ in the sense that $V_q(q, \theta) \in [p \mathbf{1}_{q \leq \theta}, (p + \delta_p) \mathbf{1}_{q \leq \theta} + \delta_q]$ for some $\delta_q, \delta_p > 0$. For any $\varepsilon > 0$, there exist $\delta_p, \delta_q > 0$ such that: (i) the best three-part tariff yields profits within ε of optimal profits; and (ii) if beliefs are symmetric around θ^* , then the best three-part tariff (with included units at zero marginal price) yields profits at least $(p(E^*[\theta | \theta \leq \theta^*] - E[\theta | \theta \leq \theta^*])/2 - \varepsilon)$ higher than the best two-part tariff.*

PROOF:

The inequalities are satisfied if $2\delta_q p + \delta_p E[\theta] + \delta_p \delta_q \leq \varepsilon$. See Appendix.

For alternative preferences, a three-part tariff is optimal only for specific consumer and firm beliefs. For example, if $V(q, \theta) = v(q) + q\theta$, marginal costs are zero, consumers are sufficiently overconfident, and the true distribution of θ is exponential, then a three-part tariff is optimal. In otherwise the same context, however, the optimal overage rate would vary if the true distribution of θ had a varying hazard rate. (If $V_{q\theta}$ is constant and consumers believe $\theta = \theta^*$ with probability one, Proposition 2 shows that the optimal overage rate is proportional to the inverse hazard rate $(1 - F(\theta))/f(\theta)$.) Nevertheless, three-part tariffs can still be a good approximation in this context even for firm beliefs with a widely varying hazard rate. If demand is linear, beliefs are uniform, consumers estimate the mean of θ correctly, and marginal cost is zero (as in Example 1 Figure 3), it can be shown that three-part tariffs capture 82 percent or more of the virtual surplus (true surplus + fictional surplus) lost by using a two-part tariff instead of the optimal fully nonlinear tariff.¹⁷ This suggests that three-part tariffs are likely to be a good approximation for a variety of preferences beyond those close to $V(q, \theta) = p \min \{q, \theta\}$.

¹⁷ Example 1 generalizes to $V_q(q, \theta) = \alpha - \beta q + \gamma \theta$ and $\theta \sim U[\underline{\theta}, \bar{\theta}]$, with consumers who underestimate the standard deviation of θ by 100Δ percent. If $q(\underline{\theta}) > 0$, then the ratio $1 - (3\text{-part loss}/2\text{-part loss})$, reported in the final row of Table 1 for Example 1, is independent of α, β, γ , or the support of θ . For zero marginal cost, the ratio can be calculated analytically as a function solely of the overconfidence level $\Delta > 0$. Its minimum is approximately 82 percent.

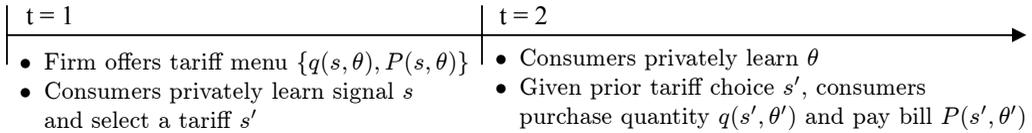


FIGURE 5. TIMELINE FOR MULTIPLE-TARIFF MODEL

H. Welfare

To evaluate welfare, I assume that the firm's prior $F(\theta)$ is correct.¹⁸ Therefore consumers' expected surplus is evaluated with respect to the firm's prior $F(\theta)$, as are expected firm profits and total surplus. Under perfect competition, welfare conclusions are straightforward. Consumers receive all the surplus generated. However, while consumers with correct priors receive the efficient allocation, overconfident consumers receive an allocation that is distorted away from first best. As a result, overconfident consumers must be worse off. This suggests that educating consumers or regulating constant marginal prices could potentially improve consumer welfare, and therefore total welfare, as firm profits are always zero.

Under monopoly, total welfare is also lower when consumers are overconfident, but in general it is ambiguous as to whether consumers or the firm are better or worse off.

LEMMA 3: *Under monopoly, if overconfident consumers overestimate the surplus created by the first-best allocation¹⁹ ($E^*[S^{FB}] \geq E[S^{FB}]$), then the firm is better off and consumers are worse off due to their overconfidence. If overconfident consumers do not overestimate the mean of θ ($E^*[\theta] \leq E[\theta]$) and $V_{\theta\theta} \geq 0$, then the firm is worse off and consumers are better off due to their overconfidence.*

PROOF:

See Appendix.

If overconfident consumers overestimate expected first-best surplus, then they are willing to pay a larger fixed fee for marginal cost pricing than a monopolist could extract from consumers with correct priors. The tables are turned if consumption value V is convex in θ and overconfident consumers do not overestimate average θ (and hence underestimate their expected utility),²⁰ because this creates bargaining power. The firm cannot extract all surplus ex ante, and to extract it ex post the firm must give away information rents as the customer is privately informed about θ in period two. This is the case in the examples discussed in Section III F, as consumers are assumed to estimate the mean of θ correctly. These consumers are strictly better off when overconfident, while the monopolist is worse off and would prefer customers to have correct priors.

¹⁸ See Footnote 10.

¹⁹ Note that, given zero marginal cost, assuming that $E^*[S^{FB}] \geq E[S^{FB}]$ is equivalent to assuming that overconfident consumers overestimate their expected value of consuming up to their satiation points.

²⁰ Given $V_{\theta\theta} \geq 0$, consumers' second-period utility $U(\theta)$ is convex in θ for any incentive-compatible allocation $q(\theta)$. Assumption A* and $E^*[\theta] \leq E[\theta]$ imply that $F(\theta)$ RSOSD $F^*(\theta)$. (RSOSD is defined in Section IV A.) Convexity and RSOSD imply $E^*[U(\theta)] \leq E[U(\theta)]$.

IV. Multiple-Tariff Menus

The model presented in Section III assumes that consumers have homogeneous priors *ex ante*, and therefore firms offer only a single tariff. In this section, I examine the case where consumer beliefs are initially heterogeneous and firms offer a menu of contracts.

Rather than assuming that *ex ante* all consumers have homogeneous prior $F^*(\theta)$, assume instead that each consumer receives a private signal $s \sim G(s)$ prior to choosing a tariff. The signal s does not enter payoffs directly, but is informative about $\theta \sim F_s^*(\theta)$. At $t = 1$, a firm (or multiple firms) will offer a menu of tariffs $\{q(s, \theta), P(s, \theta)\}$ from which consumers will choose based on their signal s . Then, at $t = 2$, consumers privately learn θ , and choose a quantity based on both θ and their prior choice of tariff (Figure 5). Preferences are identical to those in the single-tariff model, except that in the monopoly case I make the stronger restriction $V_{qq\theta} = 0$ in place of equation (1).

There are two natural dimensions of heterogeneity to explore. First, consumers could all be similarly overconfident, but vary in their average demand. Second, consumers could all have the same true distribution of demand, but vary in their degree of overconfidence. (The latter scenario could arise because experienced consumers overcome overconfidence via learning, but new customers continually enter the market.) Extending the model in this way requires separate analysis for monopoly and perfect competition market conditions.

A. Monopolist's Multiple-Tariff Menu

The monopolist's multiple-tariff problem is formally stated and analyzed in Web Appendix C for two first-period signals $s \in \{L, H\}$ (see Grubb (2008) for a continuum of first-period signals). For this analysis, I replace equation (1) by the stricter assumption $V_{qq\theta} = 0$. I consider signal spaces which order consumer beliefs $F_s^*(\theta)$ either by FOSD, or a more general reverse second-order stochastic dominance (RSOSD), defined below.²¹ This is sufficiently general to capture heterogeneity either in average demand or in degree of overconfidence.

DEFINITION 1: S is ordered by RSOSD²² if $\int_{\bar{\theta}}^{\bar{\theta}} F_H^*(x) dx \leq \int_{\bar{\theta}}^{\bar{\theta}} F_L^*(x) dx, \forall \theta$, and $\forall H \geq L$.

If there are only two first-period signals, $s \in \{L, H\}$, a monopolist will offer a menu of two tariffs $\{q_L(\theta), P_L(\theta)\}$ and $\{q_H(\theta), P_H(\theta)\}$. Beyond the familiar nonnegativity, satiation, and monotonicity constraints, substituting incentive and participation constraints in place of payments leaves only the upward incentive constraint that type L should not want to deviate and choose tariff H . In their analysis of optimal refund contracts, Courty and Li (2000) solve a related problem by first relaxing the upward incentive constraint, and then checking that it is satisfied using a sufficient (but not necessary) monotonicity condition. Within the current model, if $F_H^*(\theta)$ FOSD $F_L^*(\theta)$, then the analogous monotonicity condition that is sufficient (but not necessary) for upward incentive compatibility is $q_L(\theta) \leq q_H(\theta)$ for all θ . As this sufficient condition can easily fail for overconfident consumers,²³ a more productive approach in this context is to directly incorporate

²¹ The assumed ordering applies to the consumers' beliefs F^* , rather than the firm's beliefs F . The ordering is required to make incentive constraints tractable, and these depend on consumer beliefs.

²² This is reverse SOD rather than SOD because I integrate from the top down rather than the bottom up. Thus, higher s corresponds to higher spread and higher mean rather than lower spread and higher mean. Increasing s could thus be thought of as the combination of a mean preserving spread and a first-order stochastic dominant shift. Given that F_L crosses F_H once from below, F_H RSOSD F_L if and only if $\mu_H \geq \mu_L$.

²³ Given $F_H^*(\theta)$ FOSD $F_L^*(\theta)$, the binding downward incentive constraint is sufficient for upward incentive compatibility if $q_L(\theta) \leq q_H(\theta)$. Overconfidence can generate greater downward distortions on the H contract than on the L contract. This is because (excluding screening effects) extreme overconfidence drives overage rates upward toward *ex post* monopoly prices (Web Appendix A). These are higher (which implies greater downward quantity distortion) for

the upward incentive constraint using optimal control techniques. Doing so leads to the following characterization of marginal prices.

PROPOSITION 4: *Given (i) $F_H^*(\theta)$ FOSD $F_L^*(\theta)$, or (ii) $V_{\theta\theta} \geq 0$ and $F_H^*(\theta)$ RSOSD $F_L^*(\theta)$: the monopoly payment functions $\hat{P}_L(q)$ and $\hat{P}_H(q)$ are continuous and piece-wise smooth functions of quantity. There may be kinks in the payment functions where marginal price increases discontinuously. These kinks occur where a monotonicity constraint binds and an interval of types “pool” at the same quantity on the given tariff. Let $\hat{\theta}(q, s)$ be the inverse of the monopoly allocation $\hat{q}_s(\theta)$ when the latter is invertible. For quantities at which there is no pooling, marginal prices are given by equations (9–10) where $\alpha = \Pr(H)$ and $\gamma \geq 0$ is the shadow price of the upward incentive constraint:*

$$(9) \quad \frac{d\hat{P}_L(q)}{dq} = \max \left\{ 0, C_q(q) + V_{q\theta}(q, \theta) \frac{F_L^*(\theta) - F_L(\theta)}{f_L(\theta)} + \frac{\gamma + \alpha}{1 - \alpha} V_{q\theta}(q, \theta) \frac{F_L^*(\theta) - F_H^*(\theta)}{f_L(\theta)} \right\},$$

where $\theta = \hat{\theta}(q, L)$

$$(10) \quad \frac{d\hat{P}_H(q)}{dq} = \max \left\{ 0, C_q(q) + V_{q\theta}(q, \theta) \frac{F_H^*(\theta) - F_H(\theta)}{f_H(\theta)} - \frac{\gamma}{\alpha} V_{q\theta}(q, \theta) \frac{F_L^*(\theta) - F_H^*(\theta)}{f_H(\theta)} \right\},$$

where $\theta = \hat{\theta}(q, H)$.

PROOF:

Given results in Web Appendix C, the proof is analogous to the proof of Proposition 2 and hence omitted.

Heterogeneity in Average Demand.—To capture heterogeneity in average demand among similarly overconfident consumers, assume that consumers’ conditional priors $F_s^*(\theta)$ cross firms’ conditional priors $F_s(\theta)$ once from below at $\theta^*(s)$, and that signals are ordered by FOSD so that $F_H^*(\theta) \leq F_L^*(\theta)$. Proposition 4 applies and equations (9–10) show that when the upward incentive constraint is not binding ($\gamma = 0$), marginal prices for the high-demand user are identical to those in the single-tariff case. On the other hand, marginal prices for the low-demand user are distorted upward from the single-tariff benchmark. This matches the standard price discrimination intuition in which there is no distortion at the top, but the allocation of lower types is distorted downward to increase rent extraction from high types. If the upward incentive-constraint binds ($\gamma > 0$), the upward distortion of tariff-L marginal prices is exacerbated while tariff-H marginal prices are distorted downward.

higher average demand. Ex ante screening will ameliorate this effect by distorting quantity downward on the L contract, but not necessarily completely. The upward incentive constraint is likely to bind when the monotonicity condition is severely violated by the relaxed ($\gamma = 0$) solution, and must bind when it is violated at all θ .

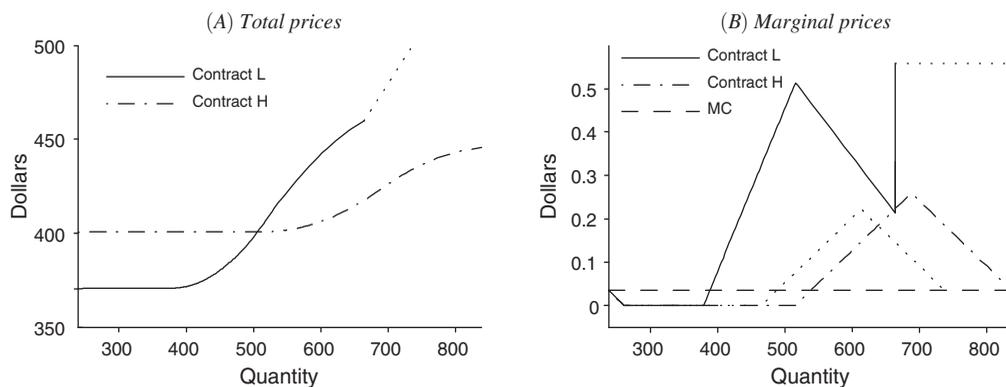


FIGURE 6. A MONOPOLIST'S TWO-CONTRACT MENU (EXAMPLE 2)

When overconfidence, measured by $|F_s^* - F_s|$, is high relative to the ex ante heterogeneity, measured by $|F_L^* - F_H^*|$, the qualitative predictions of the single-tariff model are robust. The distortions from first-period screening will tend to reduce the number of included units and increase overage rates for the low-tariff, but have the opposite effect on the high-tariff. (This fits observed cellular phone service tariff menus which involve increasing allowances and declining overage rates across tariffs.) If ex ante heterogeneity is sufficiently large relative to overconfidence, it should be expected that included units would be eliminated altogether from the low-tariff, the upward incentive constraint would not bind, and hence the high-tariff would set marginal price equal to the single-tariff benchmark.

EXAMPLE 2: I extend Example 1 to the case of two ex ante signals. Half of consumers, who receive signal *L* ex ante, have the same true and perceived distributions of demand as in the low-overconfidence case ($\Delta = 0.25$) of Example 1: F_L is $U[-0.5, 0.5]$ and F_L^* is $U[-0.375, 0.375]$. The other half of consumers, who receive signal *H* ex ante, have higher demand in a FOSD sense: F_H is $U[-0.5, 0.7]$ and F_H^* is $U[-0.35, 0.55]$. Both low and high average demand consumers estimate the mean of θ correctly but underestimate the standard deviation of θ by 25 percent. Marginal cost is \$0.035.

Figure 6 depicts a monopolist's optimal two-contract menu for Example 2. Panel A shows total prices for contracts *L* and *H*, while panel B shows marginal cost and marginal prices for contracts *L* and *H*. In this example, upward incentive constraints do not bind ($\gamma = 0$), so screening does not distort contract *H* away from the single-contract benchmark.²⁴ Ex ante screening does distort contract *L*. Panel B shows as a dotted line marginal prices that would be charged on contract *L* if there were no distortions due to screening, which coincides with the single-contract benchmark shown in panel C of Figure 4. By comparison, the solid line in panel B shows that optimal marginal prices for contract *L* are distorted upward. Due to the distortion, contract *L* provides roughly 85 fewer "free" units, and a maximum marginal price of \$0.56 rather than only \$0.22.

Despite the screening distortion, contracts *L* and *H* display key qualitative features of observed three-part tariff menus: increasing fixed fees for increasing quantities of "free" units followed

²⁴ Upward incentive compatibility is not binding, despite the fact that optimal marginal prices of the two contracts briefly cross, and hence allocations are not monotonically increasing from contract *L* to *H*.

by steep marginal charges. Moreover, unlike the single-contract benchmark, marginal price is strictly positive for the last unit sold on contract L . (Note that no consumption occurs on the dotted portion of contract L in equilibrium, and the model places only a lower bound on prices in this region. The dotted extension represents the minimum constant marginal price extension of the contract. Marginal price does not fall to marginal cost because the supports of the type distributions for low and high average demand groups do not coincide.)

Heterogeneity in Degree of Overconfidence.—To capture heterogeneity in degree of overconfidence (rather than average demand), assume that the firm's conditional priors are independent of s ($F_s(\theta) = F(\theta) \forall s$), high types have correct beliefs ($F_H^*(\theta) = F(\theta)$), and low types are overconfident ($F_L^*(\theta) = F^*(\theta)$ crosses $F(\theta)$ once from below). If overconfident consumers do not overestimate the mean of θ , so that $F_H^*(\theta)$ RSOSD $F_L^*(\theta)$, and in addition $V_{\theta\theta} \geq 0$, then Proposition 4 applies.²⁵ Equations (9–10) reduce to equations (11–12) and show that screening between overconfident and correct-prior consumers exacerbates the distortion in marginal prices for overconfident consumers. When the upward incentive constraint is not binding ($\gamma = 0$), marginal price equals marginal cost for the correct-prior consumer. On the other hand, if the upward incentive constraint binds ($\gamma > 0$), then correct-prior consumers will pay above marginal cost for initial quantities and below marginal cost for higher quantities (as if they were underconfident):

$$(11) \quad \frac{d\hat{P}_L(q)}{dq} = \max \left\{ 0, C_q(q) + \frac{1 + \gamma}{1 - \alpha} V_{q\theta}(q, \theta) \frac{F^*(\theta) - F(\theta)}{f(\theta)} \right\},$$

$$(12) \quad \frac{d\hat{P}_H(q)}{dq} = \max \left\{ 0, C_q(q) - \frac{\gamma}{\alpha} V_{q\theta}(q, \theta) \frac{F^*(\theta) - F(\theta)}{f(\theta)} \right\}.$$

Note that Lemma 3 extends to this setting, and overconfident consumers are better off than they would be if all consumers had correct beliefs. At the same time, the fraction of consumers with correct priors must be weakly better off than their overconfident counterparts, because correct beliefs lead to better decisions. Thus, the presence of overconfident types in the marketplace improves the outcome for both types.

B. Perfectly Competitive Multiple-Tariff Menu

If firms believe that all consumers share the same true distribution of θ , while consumers have heterogeneous beliefs (for instance because they differ in severity of overconfidence), then screening between consumers with different beliefs does not distort prices. Consumers who receive signal s will be offered and will choose the tariff described by the single-tariff model as if they were the only type in the market.²⁶

²⁵ If overconfident consumers overestimate the mean of θ , then Proposition 4 does not apply. In this case, solving the monopolist's problem is more difficult since it is not clear which first-period incentive and participation constraints will bind. However, there is one simple benchmark case where neither upward nor downward incentive constraints bind. If overconfident types correctly estimate the surplus generated by the first-best allocation ($E^*[S^{FB}] = E[S^{FB}]$), then correct-prior and overconfident types will be offered the same monopoly tariffs as if they were each the only type in the market: the standard tariff $\{q^{FB}(\theta), C(q^{FB}(\theta)) + E[S^{FB}]\}$ and the overconfident tariff $\{\hat{q}(\theta), P^M(\theta)\}$, respectively. (When marginal costs are zero, the benchmark $E^*[S^{FB}] = E[S^{FB}]$ implies consumers correctly estimate their expected value of consuming up to their satiation points.) For a proof, see Appendix.

²⁶ Note that a firm expects the same profit on a given tariff from all participating consumers who the firm believes have the same type distribution $F(\theta)$. Thus the set of zero-profit tariffs is the same regardless of whether a firm is

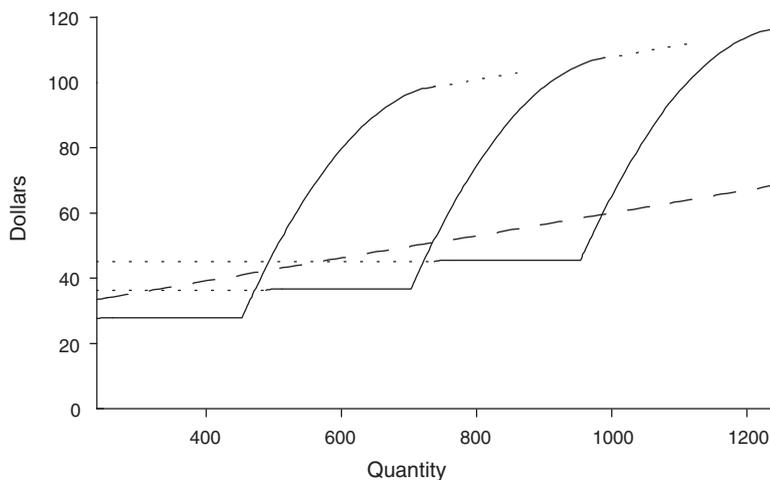


FIGURE 7. A COMPETITIVE THREE-CONTRACT MENU

Notes: Solid portions of the tariffs are uniquely optimal. Dashed portions of the tariffs are illustrative extensions where no consumption takes place. The straight line shows total costs.

A general characterization of the equilibrium tariff menu under perfect competition when consumers differ both in their beliefs and in their true distributions of demand is a difficult problem left for future research. Unlike the monopoly case, it is insufficient to solve a single constrained maximization problem. In fact, there may be no set of tariffs that yields nonnegative profits such that entry is not profitable. In other words, if firms simultaneously set prices, there may not be a pure strategy equilibrium. This is similar to Rothschild and Stiglitz's (1970) result about competitive insurance markets.

In some cases, ex ante screening is not distortionary, and each contract on the equilibrium menu matches that in the single-tariff model. One such example is illustrated by Figure 7. This is a variation of the example shown in column 2 of Figure 4, in which marginal cost is \$0.035 and consumers are highly overconfident ($\Delta = 0.75$). The variation is that consumers receive one of three signals ex ante, low, medium, or high, which correspond to future θ being distributed uniformly over the interval $[-\frac{1}{2}, \frac{1}{2}]$, $[0, 1]$, or $[\frac{1}{2}, \frac{3}{2}]$, respectively. This example yields a tariff menu qualitatively similar to cellular phone service tariff menus. Moreover, the predicted usage distributions of customers on each tariff are ordered by strict first-order stochastic dominance, which matches actual usage patterns described in Section VI.

V. Potential Alternatives

There are several potential alternatives to the model of overconfidence which are worth considering. First, existing literature considers the implications of a number of biases for optimal nonlinear pricing. These include a flat-rate bias as well as biases related to systematically underestimating usage. Moreover, by considering alternative type distributions in the context of the

serving consumers with one prior or another. The competitive single-tariff solution is the tariff in this set that maximizes the perceived expected utility of its target customer. Hence, consumers cannot believe themselves to be strictly better off under another offered tariff.

multiple-tariff monopoly model of the previous section, I am able to develop an alternative explanation of three-part tariffs based on price discrimination with common priors. I explore each of these three alternatives below.

A. Flat-Rate Bias

Several authors have documented a “flat-rate bias.” This term refers to a tendency of consumers to choose a flat-rate tariff despite the availability of a metered tariff, which would be cheaper given their usage levels (Kenneth E. Train 1991). Lambrecht and Skiera (2006) provide an overview of the work on flat-rate bias, document the bias among Internet service customers in Germany, and identify three significant causes: risk aversion, demand overestimation, and the “taxi-meter” effect (that prices directly enter consumer preferences: consumers derive less pleasure from units of consumption that accrue marginal charges, than those that are prepaid with a fixed fee).

Although the existence of a flat-rate bias may influence the terms of three-part tariffs, on its own it does not provide a complete explanation for their use. It is true that three-part tariffs are locally flat over the allowance of included units, but globally three-part tariffs are not flat. In fact, for the 1,519 cellular customers in my sample who chose a three-part tariff, overages occur on 19 percent of bills. Conditional on occurrence, this leads to average overage charges more than twice the average monthly fixed fee (see Section VI).

If consumers are risk averse to such variability in their monthly bill, the use of three-part tariffs with steep overage charges is more surprising, rather than less surprising. Similarly, a tendency of consumers to overestimate their future demand does not explain three-part tariff pricing. Proposition 2 was derived without any assumption on the relationship between consumer and firm priors; hence it can be used to analyze biases other than overconfidence. Corollary 2 confirms that demand overestimation conflicts with steep overage rates observed on three-part tariffs.

COROLLARY 2: *If consumers overestimate their future demand in an FOSD sense ($F^*(\theta) \leq F(\theta)$), then marginal price is between zero and marginal cost at all quantities.*

PROOF:

Follows from Proposition 2.

Although flat-rate bias does not appear to explain three-part tariff pricing on its own, consumers are likely to exhibit flat-rate bias and be overconfident at the same time. Overconfident consumers underestimate the likelihood of making overages, and in the extreme could believe that they will always stay within the flat portion of their chosen tariff. Given such beliefs, a flat-rate bias due to either the taxi-meter effect or risk aversion would reinforce the attractiveness of three-part tariffs. Hence, overconfidence and flat-rate bias can be viewed as complementary explanations for three-part tariff pricing.

B. Underestimates

Gabaix and Laibson (2006) examine optimal pricing of goods with add-ons (e.g., printers and printer cartridges) when some consumers are myopic or are unaware of the need to purchase the add-on. When sophisticated consumers can substitute away from the add-on through advanced planning, myopic consumers subsidize low prices of the primary good by paying high add-on

fees.²⁷ In related work, DellaVigna and Malmendier (2004) examine optimal two-part tariff pricing when consumers are quasi-hyperbolic discounters. They show that marginal prices should be set below marginal cost for investment goods (such as health clubs) and above marginal cost for leisure goods (such as cellular phone service). DellaVigna and Malmendier (2004) mention that this may explain why cell phone tariffs include marginal prices above marginal cost, but this theory does not explain why marginal prices are initially zero. Oren Bar-Gill (2006) draws similar conclusions when consumers underestimate their future usage.

Aside from the different welfare implications, in the context of the model in this paper, both naive beta-delta discounters and consumers unaware of future add-on purchases are essentially consumers who systematically underestimate their demand at the time of contracting. By Corollary 3, marginal price would therefore be predicted to be above marginal cost for all q .

COROLLARY 3: *If consumers underestimate their future demand in an FOSD sense ($F^*(\theta) \geq F(\theta)$), then marginal price is weakly greater than marginal cost for all quantities. Moreover, where the FOSD is strict, marginal price will be strictly greater than marginal cost.*

PROOF:

Follows from Proposition 2.

The assumption in this paper that consumers are overconfident (A^*) implies that consumers underestimate demand conditional on it being high ($\theta > \theta^*$), but overestimate demand conditional on it being low ($\theta < \theta^*$). The underestimation of demand above θ^* drives marginal price above marginal cost at high quantities, as would naive beta-delta discounting. It is the overestimation of demand below θ^* that drives the region of zero marginal price at low quantities. Thus, the balanced over- and underestimation of demand captured by overconfidence is necessary for the result.

C. Price Discrimination with Common Priors

Under perfect competition, common priors yield marginal cost pricing,²⁸ which cannot explain observed tariffs. Determining whether price discrimination with common priors can explain observed tariffs under monopoly or imperfect competition is not a trivial problem, however. Considering the multiple-tariff monopoly model discussed in Section IV for the special case of common priors ($F^*(\theta|s) = F(\theta|s)$) provides useful insight (see Grubb 2008).

First, if the distribution of demand is increasing in a first-order stochastic dominance (FOSD) sense, marginal price should always be above marginal cost and consumption distorted downward for all but those with the highest average demand. Given such a type distribution, price discrimination with common priors does not explain observed tariffs.

Second, given low marginal costs and free disposal, price discrimination with common priors could predict tariff menus qualitatively similar to those observed, which couple increasing fixed fees with increasing numbers of included minutes and declining overage rates. However, to do so, a rather implausible type distribution must be assumed. In particular, with a continuum of ex ante types, consumers' conditional priors over θ should satisfy equation (13) for some cutoff $\theta^*(s)$

²⁷ Glenn Ellison (2005) discusses alternative reasons (including advertising costs) why competing firms may sell base goods below cost and add-ons at unadvertised high prices, even when consumers are unbiased. However, because Ellison (2005) assumes unit demand for the add-on, the model cannot address three-part tariff pricing.

²⁸ This is true not only in the current model, but also in more general models that maintain key features of the framework, including the contracting time-line, quasi-linear utility, and deterministic costs.

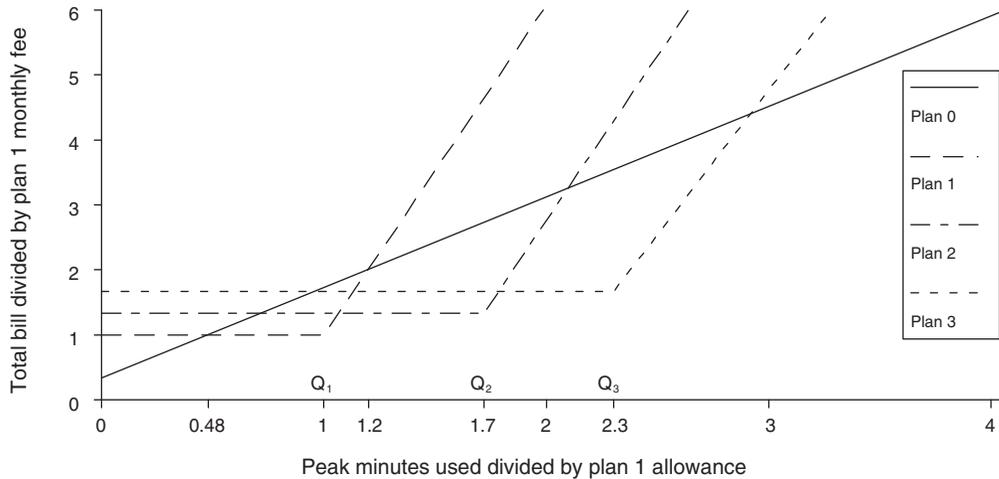


FIGURE 8. POPULAR PRICING PLANS OFFERED FALL 2002 AND FALL 2003

Note: Usage and total price are measured as a fraction of the plan 1 included allowance (Q_1) and monthly fee, respectively.

increasing in s . Alternatively, with two ex ante types, $F_L(\theta)$ must cross $F_H(\theta)$ once from below (see Proposition 4).

$$(13) \quad \frac{\partial}{\partial s} (1 - F(\theta|s)) \begin{cases} \leq 0 & \theta \leq \theta^*(s) \\ > 0 & \theta > \theta^*(s) \end{cases} .$$

To understand why this type distribution generates such pricing, consider an example with two ex ante types. The high type ($s = H$) is an undergraduate whose valuation is high on average, but is also highly variable. The undergraduate is either on campus and has a high demand, or is away on break and has a low demand. The low type ($s = L$) is a graduate student who consistently has a moderate demand somewhere in between these two extremes. In this case, a monopolist will find it optimal to offer the undergraduate user unlimited usage at marginal cost for a high monthly fee. The graduate student will pay a low monthly fee for small or zero marginal charges below marginal cost at low quantities, followed by high marginal charges above marginal cost at high quantities. The high marginal charges at high quantities have little impact on either an undergraduate on break or a graduate student, but make the graduate student tariff much less attractive to an undergraduate on campus. The initial small or zero marginal charges are attractive to the graduate student, and allow a higher monthly fee to be charged on the graduate tariff. This trade-off is a wash for an undergraduate on campus, but is unattractive to an undergraduate on break. Together, both distortions of the graduate student tariff away from marginal cost pricing increase the surplus that can be extracted from an undergraduate ex ante.

For two tariffs with $Q_1 < Q_2$ included minutes, marginal prices are zero on both tariffs for $q \in (0, Q_1)$. Thus, assumptions about the distribution of demand for consumers on each plan map directly into conclusions about distributions of consumption up to Q_1 . A type distribution described by equation (13) therefore requires that consumers selecting a tariff with $Q_2 > Q_1$ included minutes would be more likely to consume strictly less than Q_1 minutes than would consumers who actually selected the tariff with Q_1 included minutes.²⁹ More specifically, it

²⁹ Consumers who realized $\theta \leq \theta^*(s)$ would consume weakly below their included limit $Q = \hat{q}(s, \theta^*(s))$, and consumers who realized $\theta > \theta^*(s)$ might make overages.

TABLE 2—OVERAGES ARE LARGE AND FREQUENT
(Plans 1–3, fall 2002 and 2003 menus)

	Observations		Usage [†]	
	<i>n</i>	<i>n/N</i>	Mean	Standard deviation
Under allowance	6,097	83 percent	0.47	0.27
Over allowance	1,274	17 percent	1.43	0.44
Total	7,371	100 percent	0.63	0.47

[†]Measured as a fraction of included plan allowance.

requires that the cumulative usage distribution of plan 1 customers be below that of plan 2 customers, for all $q < Q_1$: $H(q|s_1) \leq H(q|s_2)$. As shown in the following section, this is not consistent with observed consumer behavior. As a result, the common-prior model does not appear to explain observed cellular phone service tariff menus.

VI. Empirical Analysis

I have obtained billing data for 2,332 student accounts managed by a major US university for a national US cellular phone service provider.³⁰ The data span 40 of the 41 months February 2002 through June 2005 (December 2002 is missing), and include 32,852 individual bills. Within the sample, there are more than 50 distinct tariffs from more than 10 menus. These include national calling plans, local calling plans, and a two-part tariff, which all vary over time.

For my primary analysis, I focus on two similar menus with the most usage data. These are the set of local plans offered to students in the fall of 2002 and the fall of 2003 (Figure 8). Within these menus I look at the four most popular plans. These include three-part tariffs with the smallest, second smallest, and third smallest monthly fixed fees and allowances, which I will refer to as plans 1, 2, and 3, respectively. While over 55 percent of students who signed up for a new tariff in the fall of 2002 or 2003 chose either plan 1, 2, or 3, an important alternative was a two-part tariff, which I call plan 0. Plan 0 has a small monthly fixed fee and a constant per-minute charge below the overage rates of plans 1–3.³¹

³⁰ Students received an itemized phone bill, mailed by default to their campus residence, which was separate from their university tuition bill. The sample of students is undoubtedly different from the entire cellular phone service customer base. However, a pricing manager from one of the top US cellular phone service providers who kindly read through an earlier draft of this paper made the unsolicited comment that the empirical findings were highly consistent with their own internal analysis of much larger and representative customer samples.

³¹ Most students sign up for cellular phone service at the beginning of the academic year, which is why the fall menus are most relevant. The fall 2002 menu was offered September 2002 through March 2003. Almost the same menu was offered again September 2003 through December 2003, although minute allowances increased by 1–2 percent, overage rates increased by 0–15 percent, and free nights and weekends were eliminated for plan 0. Since plans 1, 2, and 3 change only slightly between 2002 and 2003, I analyze usage patterns pooling data from the two menus. Analyzing usage patterns separately for 2002 and 2003 yields qualitatively similar results. In fact, a larger price change was that plan 1 had 25 percent fewer included minutes for a short time after its introduction in the fall of 2002, before being increased to the level shown in Figure 8 for both new and existing customers. Excluding these initial plan 1 bills gives qualitatively similar results for all of my analysis. Plan 0 was not offered to the general public, but only to the students who received service through the university. In the United States it is typical for such two-part tariffs to be included in corporate rate packages, but not be offered to the general public. Students received additional benefits including up to 15 percent additional included minutes on plans, and a required service commitment of only 3 months rather than 12 months.

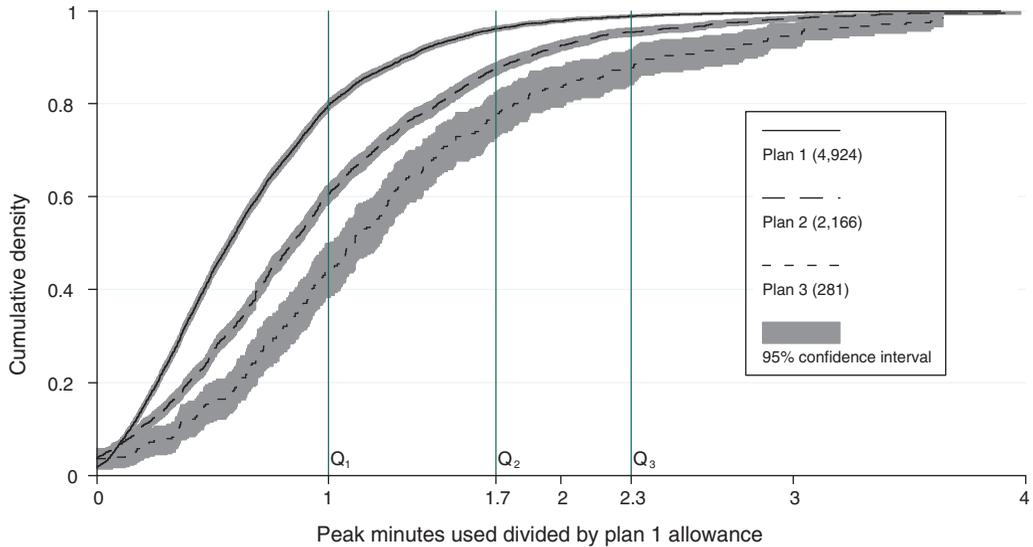


FIGURE 9. CUMULATIVE USAGE DISTRIBUTIONS FOR PLANS 1, 2, AND 3 (FALL 2002 AND 2003 MENUS)

Notes: Usage is measured as a fraction of the plan 1 allowance, (Q_1). Vertical lines are drawn at all three plan allowances. Confidence intervals are shaded in gray.

The overconfidence and price discrimination explanations of three-part tariff pricing can be distinguished by comparing customer usage patterns across different three-part tariffs on the same menu. Figure 9 plots the cumulative usage distributions $\hat{H}(q|plan)$ and their 95 percent confidence intervals³² for customers on plans 1, 2, and 3. Prorated bills for incomplete months of service are excluded, as are bills with missing usage information. In total, the distribution plotted for plan 1 is based on 4,924 bills of 498 customers, while plan 2 is based on 2,166 bills of 210 customers, and plan 3 is based on 281 bills of 31 customers.

Figure 9 shows that the three usage distributions are statistically indistinguishable at the very bottom and the very top, but everywhere else the distributions are consistent with a strict FOSD ordering. Formal pair-wise tests of first-order stochastic dominance among the three distributions provide limited additional insight.³³ It is clear from the figure, however, that usage patterns are inconsistent with the assumption driving the common-prior alternative.

It is not true that $\hat{H}(q|plan1) \leq \hat{H}(q|plan2)$ for $q < Q_1$. Plan 2 customers are not “undergraduate” types who actually consume less than Q_1 minutes more frequently than plan 1 customers. Rather, plan 2 customers use less than Q_1 minutes only 60 percent of the time, while plan 1

³² If $\hat{H}(q)$ denotes the sample cumulative density function (CDF) for N observations, a 95 percent confidence interval is calculated point-wise as $\hat{H}(q) \pm 1.96 \sqrt{(1 - \hat{H}(q)) \hat{H}(q)/N}$. This is because, for large N , $\hat{H}(q)$ is approximately normal with mean of the true CDF $H(q)$ and variance $(1 - H(q))H(q)/N$.

³³ Garry F. Barrett and Stephen G. Donald’s (2003) test fails to reject the null hypothesis of FOSD for each pair at any reasonable significance level. Yet, because the distributions are statistically indistinguishable at the top and bottom, the KRS test Yiu Kuen Tse and Xibin Zhang (2004) describe, which is based on Amarjot Kaur, B. L. S. Prakasa Rao, and Harshinder Singh (1994), fails to reject the complementary null hypothesis for each pair at a 10 percent significance level. The DD test that Tse and Zhang (2004) describe, which is based on Russell Davidson and Jean-Yves Duclos (2000), rejects the null hypothesis of distribution equality at a 1 percent significance level and accepts the first alternative hypothesis that the distributions have an FOSD ordering. (This test was based on 20 points equally spaced in the range of the plan 1 usage distribution, using a critical value from Michael R. Stoline and Hans K. Ury (1979).)

TABLE 3—FREQUENCY AND SIZE OF EX POST “MISTAKES”
(Fall 2002 menu)

	Plan 0 customers	Plan 1 customers	Plan 2 customers
Customers	393 (62 percent)	92 (15 percent)	124 (20 percent)
Bills	5,495	893	1,185
Alternative considered	Plan 1, 2, or 3	Plan 0	Plan 0
Alternative lower cost ex post	5 percent	65 percent	49 percent
Conditional average saving [†]	11 percent*	42 percent**	37 percent**
Unconditional average saving [†]	NA	20 percent**	1 percent

[†]Average per month, as a percentage of plan 1 monthly fixed fee.

**09 percent confidence.

*90 percent confidence.

TABLE 4—UNDERUSAGE VERSUS OVERUSAGE FOR CUSTOMERS WHO COULD HAVE SAVED ON PLAN 0

	Plan 1 customers (60)		Plan 2 customers (61)	
	Bills (523)	Potential saving [†]	Bills (578)	Potential saving [†]
Underusage	56 percent	20 percent	59 percent	30 percent
Intermediate	28 percent	(8 percent)	33 percent	(16 percent)
Overusage	16 percent	30 percent	9 percent	23 percent
Total	100 percent	42 percent	100 percent	37 percent

[†]Average per month as a percentage of plan 1 monthly fixed fee.

customers use less than Q_1 minutes 80 percent of the time. Comparisons with usage by plan 3 customers are similar. Therefore, in contrast to the overconfidence explanation, the alternative price discrimination model cannot simultaneously explain both observed pricing and observed usage patterns.

One might be concerned that the model of overconfidence is off the mark if one believes that customers only rarely exceed their included allowances. It is reasonable to hypothesize that typical overage rates of 35 to 45 cents are designed to be prohibitive outside of emergency situations. The model of overconfidence presented in this paper, however, explicitly incorporates the idea that many consumers will be surprised by higher demand than expected and use more than the included number of minutes.

Figure 9 and Table 2 clearly show that overages are an important feature of customer behavior. While 83 percent of the time customers on plans 1–3 do not exceed their allowance, using only half of included minutes on average, the other 17 percent of the time they exceed their allowance by an average of 43 percent. Moreover, overages are an important source of firm revenue. Within the entire dataset, there are 18,906 individual bills from 1,519 unique customers who are on a tariff with a strictly positive number of included minutes. Within this sample, 19 percent of bills contain overages. Moreover, the average overage charge is 42 percent of the average monthly fixed fee (231 percent conditional on an overage occurring), and represents 22 percent of average revenues (excluding taxes). In this regard, the model presented in this paper is consistent with customer behavior.

Consumer tariff choices indirectly reveal something about consumers' initial expectations for future usage. Comparing initial tariff choice with subsequent usage is therefore informative about consumer overconfidence. Plans 0–3 on the fall 2002 menu are identical in all dimensions, other than peak usage pricing described by Figure 8, and in particular all offered free night and weekend

calling. In fall 2003, plan 0 no longer offered free off-peak calling, and is therefore less comparable to the three-part tariffs. As a result, I focus on the fall 2002 menu for the following analysis.

Plans 1 and 2 are cheaper than plan 0 only for a relatively narrow range of consumption: between 48 percent and 120 percent of Q_1 for plan 1 and between 42 percent and 122 percent of Q_2 for plan 2 (Figure 8). Consumers' choice of plans 1 or 2 implies an initial belief that their consumption would likely fall within these bounds. In fact, bills of plan 1 and 2 customers fall outside these bounds, both above and below, roughly half of the time. As a result, a large fraction of consumers make ex post "mistakes," in the sense that, cumulatively over the duration of these customers' tenure in the data with a chosen plan, an alternative plan would have been lower cost for the same usage.

Table 3 shows that at least 65 percent of plan 1 customers would have saved money by initially choosing plan 0, and would have saved an average of 42 percent of the plan 1 monthly fixed fee.³⁴ In fact, had all plan 1 customers chosen plan 0 instead, they would have saved an average of 20 percent of the plan 1 monthly fixed fee. Similarly, 49 percent of plan 2 customers would have saved an average of 37 percent of the plan 1 monthly fixed fee by choosing plan 0, although had all plan 2 customers chosen plan 0, average savings would not be significantly different from zero. In contrast, only 5 percent of plan 0 customers would have saved money by choosing plans 1, 2, or 3. In all cases, customers who quit or switch plans in fewer than six months make more and larger mistakes than those who stay with their chosen plan longer. This is consistent with learning, but nevertheless mistakes are still prevalent among experienced customers.³⁵

For the plan 1 and 2 customers who could have saved money on plan 0, Table 4 shows the rate of underusage and overusage, as well as their contributions to the potential savings on plan 0. Underusage (overusage) occurs in any specific month when plan 0 would have been cheaper due to low (high) usage. On a monthly basis, the most common ex post mistake is underusage, which does not justify payment of a high fixed fee. However, the less frequent ex post error of overusage is typically a much more expensive mistake due to high overage charges. Hence when dollar weighted, both underusage and overusage are important sources of potential savings. For the 65 percent of plan 1 customers who could have saved money on plan 0, potential savings from underusage occur on 56 percent of bills and average 20 percent of the plan 1 fixed fee across all bills. Potential savings from overusage occur on only 16 percent of bills, but average 30 percent of the plan 1 fixed fee across all bills. These potential savings are partially offset by 28 percent of bills in which there is neither underusage nor overusage. Figures for plan 2 customers are similar, although dollar weighted underusage contributes slightly more to mistakes than overusage, rather than the reverse.

The prevalence and size of ex post mistakes show that customers are uncertain about their future demand when making tariff choices, and that modeling this uncertainty is critical for understanding the market. Moreover, the consistent direction of the mistakes, and in particular

³⁴ The frequency and size of mistakes are underestimated. First, plan 0 includes free in-network calling, which plans 1–3 do not. I am able to correctly calculate the counterfactual cost of plans 1–3 to a plan 0 customer, but I overestimate the counterfactual cost of plan 0 to a plan 1 or 2 customer because I cannot distinguish in-network from out-of-network calls made by plan 1–3 users. Second, I do not account for the fact that customers could alter usage if enrolled in an alternative plan, making any potential switch more attractive. Moreover, if the entire choice set of plans is considered as possible alternatives, the frequency and size of ex post mistakes is substantially higher.

³⁵ For instance, consider the half of plan 2 customers who would have saved money on plan 0. Those who switch or quit in fewer than six months would have saved 64 percent of the plan 1 fixed monthly fee, while longer-term plan 2 customers would have saved only 33 percent of the plan 1 fixed monthly fee. To avoid overstating the size of mistakes, I report bill-weighted average mistake sizes in the text and Tables 3 and 4. Since those customers who make the largest mistakes tend to switch or quit earlier than others, and therefore have fewer bills, customer-weighted average mistake sizes are larger. Moreover, overusage is relatively more important for customer-weighted average mistakes, due to the presence of customers who switch plans or quit after a handful of extremely large overages. Customer-weighted averages are reported in Web Appendix D.

the fact that on average plan 1 customers could have saved money had they all chosen plan 0, provide evidence that consumers have biased beliefs *ex ante*. Were mistakes due primarily to underusage, the consumers' bias might be systematic overestimation of demand and could cause flat-rate bias (Lambrecht and Skiera 2006). Were mistakes due primarily to overusage, the consumers' bias might be systematic underestimation of demand, consistent with naive quasi-hyperbolic discounting (DellaVigna and Malmendier 2004). Either of these two biases could be captured by Eliaz and Spiegler's (2008) model. However, the important contribution of both underusage and overusage to mistakes suggests that consumers underestimate both overusage and underusage, consistent with consumer overconfidence.

Note that, in this context, the evidence suggests ruling out only one of three causes of flat-rate bias identified by Lambrecht and Skiera (2006): systematic overestimation of demand. Either of the other two causes, the taxi-meter effect or risk aversion, likely induce flat-rate bias in addition to overconfidence. Since bills fall within the flat portion of three-part tariffs over 80 percent of the time, and overconfident consumers underestimate the frequency and size of overages, such a flat-rate bias may reinforce the attractiveness of three-part tariffs and complement the overconfidence explanation for pricing.

VII. Conclusion

This paper has shown that, given overconfident consumers, low marginal costs, and free disposal, optimal pricing involves included units at zero marginal price followed by high marginal charges. Overconfident consumers overestimate the likelihood of consuming initial units and hence are willing to overpay for them *ex ante*. A zero marginal price for these units ensures consumers surprised by low demand do not receive a refund. Overconfident consumers also underestimate the likelihood of extremely high usage. Thus, high marginal charges at high quantities serve to extract value that consumers don't anticipate *ex ante*, but realize *ex post* when they are surprised by high demand. This provides a promising explanation for the three-part tariff menus observed in the cellular phone services market.

The theory ignores tariff complexity costs, and hence does not necessarily predict overage rates to be constant as observed. When consumers are primarily uncertain about the volume of desirable units, relative to a fairly consistent value for units that are desirable, three-part tariffs will be a good approximation for optimal pricing in the sense that they provide a large improvement over two-part tariffs relative to the remaining approximation losses. An example with linear demand and uniform beliefs suggests that this is true for other reasonable cases as well.

Empirical evidence shows that consumer usage patterns are consistent with the overconfidence explanation and, in particular, that *ex post* "mistakes" by consumers are consistent with the underlying assumption of overconfidence. Although an alternative common-prior explanation exists, it appears to be inconsistent with consumer usage patterns.

I have not modeled the possibility that over time consumer beliefs become calibrated correctly. Even if this happens in the long run, the psychology literature suggests that learning is slow (Marc Alpert and Howard Raiffa 1982; Scott Plous 1995; Fergus Bolger and Dilek Önköl Atay 2004). As a result, my characterization of a one-month contract may be a good approximation for an optimal multiple-month contract. Moreover, the multiple-tariff model of Section IV can capture the heterogeneity in overconfidence that arises when experience eliminates overconfidence for some fraction of the population, but new consumers who are overconfident continue to enter the market. In this case, firms would be expected to offer three-part tariffs to overconfident consumers, but less convex tariffs to sophisticated consumers. It may be that AT&T's "rollover" contracts, which carry forward unused minutes from the included allowance from one month to

the next, are specifically targeting such sophisticated consumers who realize that their demand varies month to month.

The model presented here is broadly applicable beyond cellular services, and is potentially relevant in any market in which consumers commit to a contract while they are uncertain about their eventual demand. In particular, the model can explain the use of three-part tariffs for other communication services such as Internet access, car leasing, a range of rental services, consumer credit card debt, and an increasing number of other services.

APPENDIX

PROOF OF LEMMA 1:

- (i) Monopoly: Consider any optimal tariff $\{q(\theta), P(\theta)\}$. If there is a type θ' who is offered $q(\theta') > q^S(\theta')$, a monopolist would be weakly better off offering $q(\theta') = q^S(\theta')$ instead. A type θ' consumer's equilibrium consumption, and hence incentive and participation constraints, would be unaffected. Any other consumer θ now finds it weakly less attractive to deviate by claiming type θ' , so type θ consumer's choice remains incentive compatible. Production costs would be weakly reduced, however. This argument can be repeated for any θ' ; hence the tariff $\{\min\{q(\theta), q^S(\theta)\}, P(\theta)\}$ must be weakly more profitable, and therefore still optimal. Moreover, if costs are strictly increasing, lowering $q(\theta')$ to $q^S(\theta')$ for a positive measure of types θ' would strictly increase profits.
- (ii) Perfect Competition: The argument is similar. Now, however, reducing production costs relaxes the firm participation constraint, allowing the firm to reduce $P(\theta)$ by a fixed amount and improve consumers' perceived expected utility $E^*[U(\theta)]$ without affecting incentive constraints.

RESTATEMENT OF PROPOSITION 1:

Proposition 1 may be restated as follows:

- (i) If quantity payment pair $\{\hat{q}(\theta), P^C(\theta)\}$ solves the perfect competition problem defined in Section IIIB, then the pair $\{\hat{q}(\theta), P^C(\theta) + E[\Psi(\hat{q}(\theta), \theta)]\}$ solves the monopoly problem defined in that section. Conversely, if the pair $\{\hat{q}(\theta), P^M(\theta)\}$ solves the monopoly problem, then $\{\hat{q}(\theta), P^M(\theta) - E[\Psi(\hat{q}(\theta), \theta)]\}$ solves the perfect competition problem.
- (ii) If a quantity payment pair $\{q(\theta), P(\theta)\}$ solves either the monopoly or perfect competition problems stated in Section IIIB, then the pair $\{\hat{q}(\theta), P(\theta)\}$, where $\hat{q}(\theta) = \min\{q(\theta), q^S(\theta)\}$, solves the same problem. Moreover, $\hat{q}(\theta)$ maximizes expected virtual surplus $E[\Psi(q(\theta), \theta)]$ subject to monotonicity, nonnegativity, and satiation constraints, where virtual surplus is defined by equation (4), and $P(\theta)$ is given by equation (5) for monopoly or equation (6) for perfect competition. Conversely, if $\hat{q}(\theta)$ maximizes expected virtual surplus subject to monotonicity, nonnegativity, and satiation constraints, then there exists a unique pair of payments $\{P^M(\theta), P^C(\theta)\}$ such that $\{\hat{q}(\theta), P^M(\theta)\}$ solves the monopoly problem and $\{\hat{q}(\theta), P^C(\theta)\}$ solves the perfect competition problem. These payments are given by equations (5) and (6), respectively.

PROOF OF LEMMA 2:

- (i) Under maintained assumptions, the constraint set $[0, q^S(\theta)]$ is convex and compact valued, continuous, and nonempty. Further, as shown below, virtual surplus $\Psi(q, \theta)$ is

continuous and strictly quasi-concave in q . (Note that this is where I use the restriction on $V_{qq\theta}$ described by equation (1).) Therefore $q^R(\theta)$ is a continuous function. Moreover, as $\Psi(q, \theta)$ is twice continuously differentiable in q for all θ , $q^R(\theta)$ is characterized by the first-order condition $\Psi_q(q, \theta) = 0$, unless either the nonnegativity or free disposal constraints are binding.

Virtual surplus $\Psi(q, \theta)$ is strictly quasi-concave in q :

Case a: $(F(\theta) - F^*(\theta))/f(\theta) > 0$: If $q \leq q^{FB}(\theta)$ then $V_q - C_q \geq 0$ and so Ψ is strictly increasing in q :

$$\Psi_q(q, \theta) = \underbrace{V_q(q, \theta) - C_q(q)}_{\geq 0 (q \leq q^{FB}(\theta))} + \underbrace{V_{q\theta}(q, \theta)}_{> 0} \underbrace{\frac{F(\theta) - F^*(\theta)}{f(\theta)}}_{> 0 \text{ (case a)}} > 0.$$

If $q \in (q^{FB}(\theta), q^S(\theta))$, then by equation (1) $V_{qq\theta} \leq 0$, and Ψ is strictly concave:

$$\Psi_{qq}(q, \theta) = \underbrace{V_{qq}(q, \theta) - C_{qq}(q)}_{< 0} + \underbrace{V_{qq\theta}(q, \theta)}_{\leq 0 (q > q^{FB})} \underbrace{\frac{F(\theta) - F^*(\theta)}{f(\theta)}}_{> 0 \text{ (case a)}} > 0.$$

Since Ψ is continuous, this is sufficient for strict quasi-concavity on $[0, q^S(\theta)]$.

Case b: $(F(\theta) - F^*(\theta))/f(\theta) < 0$. Similarly, if $q \geq q^{FB}$ then $V_q - C_q \leq 0$ and so Ψ is strictly decreasing in q . If $0 < q < q^{FB}(\theta)$, then by equation (1) $V_{qq\theta} \geq 0$, and Ψ is strictly concave. Since Ψ is continuous, this is sufficient for strict quasi-concavity on $[0, q^S(\theta)]$.

Case c: $(F(\theta) - F^*(\theta))/f(\theta) = 0$. Virtual surplus equals true surplus, which is strictly concave and hence also strictly quasi-concave.

- (ii) It follows from the application of standard results in optimal control theory (Atle Seierstad and Knut Sydsæter 1987) and the Kuhn-Tucker theorem. Note that because the virtual surplus function is strictly quasi-concave, but not necessarily strictly concave, optimal control results give necessary rather than sufficient conditions for the optimal allocation. (See Web Appendix B for additional discussion.)

PROOF OF PROPOSITION 2:

- (i) Differentiating equation (6) for $P(\theta)$ and making a change of variables yields an expression for marginal price that is valid at nonpooling quantities where $\hat{q}(\theta)$ is invertible: $(dP(q)/dq) = V_q(q, \hat{\theta}(q))$.
- (ii) At any pooling quantity q , by assumption $V_{q\theta} > 0$, marginal price must increase discontinuously, as above and below q marginal price is given by $V_q(q, \inf\{\theta: \hat{q}(\theta) = q\})$ and $V_q(q, \sup\{\theta: \hat{q}(\theta) = q\})$, respectively.

- (iii) At nonpooling quantities, Lemma 2 implies that either $\hat{q}(\theta)$ satisfies the first order condition $\Psi_q(\hat{q}(\theta), \theta) = 0$, or satiation binds and marginal price is zero since $V_q(q^S(\theta), \theta) = 0$. In the former case, the first-order condition implies marginal price is:

$$V_q(q, \theta) = C_q(q) + V_{q\theta}(q, \theta) \frac{F^*(\theta) - F(\theta)}{f(\theta)}.$$

When $[C_q(q) + V_{q\theta}(q, \hat{\theta}(q)) (F^*(\hat{\theta}(q)) - F(\hat{\theta}(q)))/f(\hat{\theta}(q))]$ is negative, the first-order condition $\Psi_q = 0$ implies $V_q(q, \theta)$ is negative and therefore $q \geq q^S(\theta)$. This is precisely when the satiation constraint binds, ensuring marginal price to be weakly positive. Thus marginal price is equal to $[C_q(q) + V_{q\theta}(q, \hat{\theta}(q)) (F^*(\hat{\theta}(q)) - F(\hat{\theta}(q)))/f(\hat{\theta}(q))]$ whenever that quantity is positive, and zero otherwise.

- (iv) As $\hat{\theta}(q)$ is a continuous function at nonpooling quantities, marginal price is as well. Thus, payment $\hat{P}(\hat{\theta}(q))$ is continuously differentiable at nonpooling quantities. Moreover, $\hat{P}(\hat{\theta}(q))$ is continuous because incentive compatibility requires that types who pool at the same quantity pay the same price.

PROOF OF PROPOSITION 3:

- (i) The first step is to derive an upper bound for optimal profits. Let $Q \equiv \hat{q}(\theta^*)$. Then Q units are included at zero marginal price on the optimal tariff (Corollary 1). The optimal tariff earns $\int_{\theta^*}^{\bar{\theta}} V(\min\{Q, q^S(\theta)\}, \theta) f^*(\theta) d\theta$ via the fixed fee for the perceived expected value of Q included units. Since $\hat{q}(\theta)$ is nondecreasing, profits from units above Q can be written as:

$$\int_{\theta^*}^{\bar{\theta}} ((V(\hat{q}(\theta), \theta) - V(Q, \theta)) f^*(\theta) + (P(\hat{q}(\theta)) - P(Q))(f(\theta) - f^*(\theta))) d\theta.$$

Now, integrating by parts, and applying local incentive compatibility $(dU(\theta)/d\theta = V_\theta(q(\theta), \theta))$ implies

$$\begin{aligned} & \int_{\theta^*}^{\bar{\theta}} (V(\hat{q}(\theta), \theta) - V(Q, \theta) - P(\hat{q}(\theta)) + P(Q))(f(\theta) - f^*(\theta)) d\theta \\ &= \int_{\theta^*}^{\bar{\theta}} (V_\theta(\hat{q}(\theta), \theta) - V_\theta(Q, \theta)) (F^*(\theta) - F(\theta)) d\theta. \end{aligned}$$

Since $F(\theta) \geq F^*(\theta)$ for all $\theta \geq \theta^*$, and $V_{q\theta} > 0$, this must be positive. Rearranging terms gives

$$\begin{aligned} & \int_{\theta^*}^{\bar{\theta}} ((V(\hat{q}(\theta), \theta) - V(Q, \theta)) f^*(\theta) + (P(\hat{q}(\theta)) - P(Q))(f(\theta) - f^*(\theta))) d\theta \\ & \leq \int_{\theta^*}^{\bar{\theta}} (V(\hat{q}(\theta), \theta) - V(Q, \theta)) f(\theta) d\theta. \end{aligned}$$

TABLE 5—MONOPOLY PAYOFFS

	$E[U]$	$E^*[U]$	$E[\Pi]$
Standard tariff $\{q^{FB}, C + E[S^{FB}]\}$	0	$E[\Psi^{FB}] - E[S^{FB}]$	$E[S^{FB}]$
Overconfident tariff $\{\hat{q}, P^M\}$	$E[S^*] - E[\hat{\Psi}]$	0	$E[\hat{\Psi}]$

Given the satiation constraint and fixed cost FC , this implies that an upper bound on profits is

$$\Pi^{\text{optimal}} \leq \int_{\underline{\theta}}^{\bar{\theta}} V(\min\{Q, q^S(\theta)\}, \theta) f^*(\theta) \, d\theta + \int_{\theta^*}^{\bar{\theta}} (V(q^S(\theta), \theta) - V(Q, \theta)) f(\theta) \, d\theta - FC.$$

Moreover, since preferences are close to $p \min\{q, \theta\}$:

$$\Pi^{\text{optimal}} \leq \int_{\underline{\theta}}^{\bar{\theta}} V(\min\{Q, q^S(\theta)\}, \theta) f^*(\theta) \, d\theta + (p + \delta_p) \int_{Q - \delta_q}^{\bar{\theta}} (\theta + \delta_q - Q) f(\theta) \, d\theta - FC.$$

The second step is to compare the preceding upper bound on optimal tariff profits, to minimum profits on a (possibly suboptimal) three-part tariff, which also includes Q units at zero marginal price and charges an overage rate p . Like the optimal tariff, this three-part tariff earns $\int_{\underline{\theta}}^{\bar{\theta}} V(\min\{Q, q^S(\theta)\}, \theta) f^*(\theta) \, d\theta$ for the first Q units. In addition, it earns at least $\int_Q^{\bar{\theta}} p(\theta - Q) f(\theta) d\theta$ from units above Q . Therefore, lost profits due to using a three-part approximation are at most $\delta_q p + \delta_p E[\theta] + \delta_q \delta_p$.

- (ii) Let p^* be the optimal two-part tariff marginal charge. I will derive a lower bound for the profit difference between the optimal two-part tariff, and a (possibly suboptimal) three-part tariff with $Q = \theta^*$ included units and overage rate equal to $\max\{p, p^*\}$.

Starting from the optimal two-part tariff, I analyze the shift to the three-part tariff in three stages. First, if $p^* > p$, reduce marginal price everywhere to $\min\{p, p^*\}$. At worst, this reduces marginal revenues and profits by $\delta_p E[\theta] + \delta_p \delta_q$. Second, increase marginal price from $\min\{p, p^*\}$ to p for units above $Q = \theta^*$. Since all customers still buy at least θ units, types above θ^* generate additional marginal fees, partially offset by reduced fixed fees, on units $q \in [\theta^*, \theta]$ that increase profits by

$$(p - \min\{p, p^*\}) \int_{\theta^*}^{\bar{\theta}} (\theta - \theta^*) (f(\theta) - f^*(\theta)) \, d\theta.$$

Third, reduce marginal price from $\min\{p, p^*\}$ to zero for units below $Q = \theta^*$. Both before and after the marginal price reduction, any type θ bought all units $q \leq \min\{\theta, \theta^*\}$. So all types generate additional fixed fees, partially offset by reduced marginal charges, on units $q \in [0, \min\{\theta, \theta^*\}]$ that increase profits by

$$\min\{p, p^*\} \int_{\underline{\theta}}^{\bar{\theta}} \min\{\theta, \theta^*\} (f^*(\theta) - f(\theta)) \, d\theta.$$

Now, both the second and third price changes additionally affect profits extracted from all types for units $q \in [\theta, \theta + \delta_q]$. Even if the two-part tariff with marginal price $\min\{p, p^*\}$ captured all the surplus on these units, first through the fixed fee ex ante and then again in marginal fees ex post, while the three-part tariff captured none, the loss would be at most $2\delta_q p$ (extraction of value above p is unaffected by the second and third price changes). Putting everything together, profits increase by at least

$$\begin{aligned} \Delta \Pi &\geq (p - \min\{p, p^*\}) (1 - F(\theta^*)) (E[\theta | \theta \geq \theta^*] - E^*[\theta | \theta \geq \theta^*]) \\ &+ \min\{p, p^*\} F(\theta^*) (E^*[\theta | \theta \leq \theta^*] - E[\theta | \theta \leq \theta^*]) - 2\delta_q p - \delta_p E[\theta] - \delta_p \delta_q. \end{aligned}$$

For symmetric distributions, this implies a profit increase of at least

$$\Delta \Pi \geq \frac{P}{2} (E^*[\theta | \theta \leq \theta^*] - E[\theta | \theta \leq \theta^*]) - 2\delta_q p - \delta_p E[\theta] - \delta_p \delta_q.$$

PROOF OF LEMMA 3:

- (i) Given the characterization of \hat{q} in Proposition 1, expected virtual surplus must be weakly higher under the equilibrium allocation than under the first-best allocation: $E[\hat{\Psi}] \geq E[\Psi^{FB}]$. Moreover, under marginal cost pricing, expected profits are equal to the fixed fee regardless of the prior over θ . Thus, under the first-best allocation, the expected virtual surplus is equal to the perceived expected surplus: $E[\Psi^{FB}] = E^*[U^{FB}] + E[\Pi^{FB}] = E^*[U^{FB}] + E^*[\Pi^{FB}] = E^*[S^{FB}]$. Together, this implies that $E[\hat{\Psi}] \geq E^*[S^{FB}]$. The assumption $E^*[S^{FB}] \geq E[S^{FB}]$ therefore implies that the firm is better off: $E[\hat{\Psi}] \geq E[S^{FB}]$. This in turn implies that consumers are worse off as total welfare is lower.
- (ii) Differentiating consumer utility shows that $U(\theta)$ is convex in θ for any incentive-compatible allocation if $V_{\theta\theta} \geq 0$. This follows from local incentive compatibility, increasing differences $V_{q\theta} > 0$, and monotonicity $q_\theta(\theta) \geq 0$:

$$U_{\theta\theta}(\theta) = \underbrace{V_{q\theta}(q(\theta), \theta)}_{(+)} \underbrace{q_\theta(\theta)}_{(+ \text{ by IC})} + \underbrace{V_{\theta\theta}(q(\theta), \theta)}_{\geq 0 \text{ (assumption)}} \geq 0.$$

Convexity implies that $E[U(\theta)] \geq E^*[U(\theta)]$ if F RSOSD F^* . (For RSOSD see Definition 1. The result is analogous to the standard result that if X second-order stochastically dominates (SOSD) Y , then $E[h(X)] \geq E[h(Y)]$ for any concave utility h (Rothschild and Stiglitz 1970; Joseph Hadar and William R. Russell 1969; Hanoch and Levy 1969). The proof is similar and hence omitted.) Assumption A^* and $E^*[\theta] \leq E[\theta]$ jointly imply that F RSOSD F^* . This fact is analogous to Hanoch and Levy's (1969) Theorem 3. Coupled with the participation constraint $E^*[U(\theta)] \geq 0$, $E[U(\theta)] \geq E^*[U(\theta)]$ implies that overconfident consumers are weakly better off for being overconfident, and hence the monopolist is weakly worse off. (Note that for the first-best allocation, $E[U(\theta)] \geq E^*[U(\theta)]$ implies that $E[S^{FB}] \geq E^*[S^{FB}]$, as marginal cost pricing ensures that expected profits are independent of the perceived distribution of quantity sold. Hence, parts (i) and (ii) of the lemma are consistent.)

PROOF OF CLAIM IN FOOTNOTE 25:

Table 5 gives each party's perceived monopoly payoff under the standard and overconfident tariffs. The assumption $E^*[S^{FB}] = E[S^{FB}]$ ensures that overconfident types are indifferent between the two tariffs, because $E^*[S^{FB}] = E[\Psi^{FB}]$ (see proof of Lemma 3). Further, by Lemma 3, it guarantees that consumers with correct priors will weakly prefer the first-best tariff.

REFERENCES

- Alpert, Marc, and Howard Raiffa.** 1982. "A Progress Report on the Training of Probability Assessors." In *Judgment Under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky, 294–305. Cambridge: Cambridge University Press.
- Bar-Gill, Oren.** 2006. "Bundling and Consumer Misperception." *The University of Chicago Law Review*, 73(1): 33–61.
- Baron, David P., and David Besanko.** 1984. "Regulation and Information in a Continuing Relationship." *Information Economics and Policy*, 1(3): 267–302.
- Barrett, Garry F., and Stephen G. Donald.** 2003. "Consistent Tests for Stochastic Dominance." *Econometrica*, 71(1): 71–104.
- Bolger, Fergus, and Dilek Önköl-Atay.** 2004. "The Effects of Feedback on Judgmental Interval Predictions." *International Journal of Forecasting*, 20(1): 29–39.
- Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang.** 2007. "Projection Bias in Catalog Orders." *American Economic Review*, 97(4): 1217–49.
- Courty, Pascal, and Hao Li.** 2000. "Sequential Screening." *Review of Economic Studies*, 67(4): 697–717.
- Davidson, Russell, and Jean-Yves Duclos.** 2000. "Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality." *Econometrica*, 68(6): 1435–64.
- DellaVigna, Stefano, and Ulrike Malmendier.** 2004. "Contract Design and Self-Control: Theory and Evidence." *Quarterly Journal of Economics*, 119(2): 353–402.
- Eliaz, Kfir, and Ran Spiegler.** 2006. "Contracting with Diversely Naive Agents." *Review of Economic Studies*, 73(3): 689–714.
- Eliaz, Kfir, and Ran Spiegler.** 2008. "Consumer Optimism and Price Discrimination." *Theoretical Economics*, 3(4): 459–97.
- Ellison, Glenn.** 2005. "A Model of Add-on Pricing." *Quarterly Journal of Economics*, 120(2): 585–637.
- Esteban, Susanna, and Eiichi Miyagawa.** 2005. "Optimal Menu of Menus with Self-Control Preferences." <http://www.najecon.org/naj/cache/78482800000000455.pdf>.
- Esteban, Susanna, Eiichi Miyagawa, and Matthew Shum.** 2007. "Nonlinear Pricing with Self-Control Preferences." *Journal of Economic Theory*, 135(1): 306–38.
- Fudenberg, Drew, and Jean Tirole.** 1991. *Game Theory*. Cambridge, MA: MIT Press.
- Gabaix, Xavier, and David Laibson.** 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics*, 121(2): 505–40.
- Grubb, Michael D.** 2008. "Selling to Overconfident Consumers." <http://ssrn.com/abstract=721701>.
- Hadar, Josef, and William R. Russell.** 1969. "Rules for Ordering Uncertain Prospects." *American Economic Review*, 59(1): 25–34.
- Hanoch, Giora, and Haim Levy.** 1969. "The Efficiency Analysis of Choices Involving Risk." *Review of Economic Studies*, 36(107): 335–46.
- Jensen, Sissel.** 2006. "Implementation of Competitive Nonlinear Pricing: Tariffs with Inclusive Consumption." *Review of Economic Design*, 10(1): 9–29.
- Kaur, Amarjot, B. L. S. Prakasa Rao, and Harshinder Singh.** 1994. "Testing for Second-Order Stochastic Dominance of Two Distributions." *Econometric Theory*, 10(5): 849–66.
- Lambrecht, Anja, and Bernd Skiera.** 2006. "Paying Too Much and Being Happy About It: Existence, Causes, and Consequences of Tariff-Choice Biases." *Journal of Marketing Research*, 43(2): 212–23.
- Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips.** 1982. "Calibration of Probabilities: The State of the Art to 1980." In *Judgment Under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky, 306–34. Cambridge: Cambridge University Press.
- Loewenstein, George, Ted O'Donoghue, and Matthew Rabin.** 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics*, 118(4): 1209–48.
- Maskin, Eric, and John Riley.** 1984. "Monopoly with Incomplete Information." *RAND Journal of Economics*, 15(2): 171–96.

- Miravete, Eugenio J.** 1996. "Screening Consumers through Alternative Pricing Mechanisms." *Journal of Regulatory Economics*, 9(2): 111–32.
- Miravete, Eugenio J.** 2005. "The Welfare Performance of Sequential Pricing Mechanisms." *International Economic Review*, 46(4): 1321–60.
- Mirrlees, James A.** 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies*, 38(114): 175–208.
- Mussa, Michael, and Sherwin Rosen.** 1978. "Monopoly and Product Quality." *Journal of Economic Theory*, 18(2): 301–17.
- Oi, Walter Y.** 1971. "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly." *Quarterly Journal of Economics*, 85(1): 77–96.
- Oster, Sharon M., and Fiona M. Scott Morton.** 2005. "Behavioral Biases Meet the Market: The Case of Magazine Subscription Prices." *BE Journal of Economic Analysis and Policy: Advances in Economic Analysis and Policy*, 5(1): 1–30.
- Plous, Scott.** 1995. "A Comparison of Strategies for Reducing Interval Overconfidence in Group Judgments." *Journal of Applied Psychology*, 80(4): 443–54.
- Riordan, Michael H., and David E. M. Sappington.** 1987. "Awarding Monopoly Franchises." *American Economic Review*, 77(3): 375–87.
- Rothschild, Michael, and Joseph E. Stiglitz.** 1970. "Increasing Risk: I. A Definition." *Journal of Economic Theory*, 2(3): 225–43.
- Rothschild, Michael, and Joseph E. Stiglitz.** 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics*, 90(4): 630–49.
- Sandroni, Alvaro, and Francesco Squintani.** 2007. "Overconfidence, Insurance, and Paternalism." *American Economic Review*, 97(5): 1994–2004.
- Seierstad, Atle, and Knut Sydsaeter.** 1987. *Optimal Control Theory with Economic Applications*. Amsterdam: Elsevier Science, North-Holland.
- Stole, Lars A.** 1995. "Nonlinear Pricing and Oligopoly." *Journal of Economics and Management Strategy*, 4(4): 529–62.
- Stoline, Michael R., and Hans K. Ury.** 1979. "Tables of the Studentized Maximum Modulus Distribution and an Application to Multiple Comparisons Among Means." *Technometrics*, 21(1): 87–93.
- Train, Kenneth E.** 1991. *Optimal Regulation: The Economic Theory of Natural Monopoly*. Cambridge, MA: MIT Press.
- Tse, Yiu Kuen, and Xibin Zhang.** 2004. "A Monte Carlo Investigation of Some Tests for Stochastic Dominance." *Journal of Statistical Computation and Simulation*, 74(5): 361–78.
- Uthemann, Andreas.** 2005. "Competitive Screening of Customers with Non-Common Priors." <http://www.ucl.ac.uk/~uctpaut/research/uthemannfincp.pdf>.
- Wilson, Robert.** 1993. *Nonlinear Pricing*. Oxford: Oxford University Press.

This article has been cited by:

1. Ronald L Goettler, Karen Clay. 2011. Tariff Choice with Consumer Learning and Switching Costs. *Journal of Marketing Research* **48**:4, 633-652. [[CrossRef](#)]
2. Pu-yan Nie. 2011. A note on dynamic Stackelberg games with leaders in turn. *Nonlinear Analysis: Real World Applications* . [[CrossRef](#)]
3. X. Henry Wang, Bill Z. Yang. 2010. The sunk-cost effect and optimal two-part pricing. *Journal of Economics* **101**:2, 133-148. [[CrossRef](#)]
4. Pu-yan Nie. 2010. Dynamic discrete-time multi-leader–follower games with leaders in turn. *Computers & Mathematics with Applications* . [[CrossRef](#)]